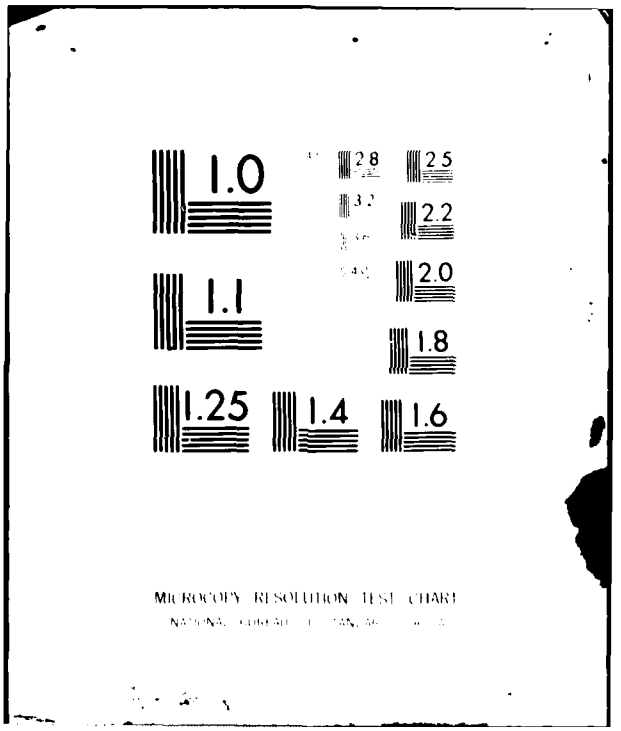
AD-A097 535 CLEMSON UNIV SC DEPT OF MATHEMATICAL SCIENCES F/G 21/1
AN EMPIRICAL MODEL BUILDING CRITERION BASED ON PREDICTION WITH --ETC(U)
AUG 80 A S KORKOTSIDES, K T WALLENIUS N00014-75-C-0451
UNCLASSIFIED N117 NL

1 OF 2


1.0
2.8
2.5
3.2
2.2
2.0
1.1
1.8
1.25
1.4
1.6
MICROCOPY RESOLUTION TEST CHART

LEVEL II

(6)

AN EMPIRICAL MODEL BUILDING CRITERION BASED ON PREDICTION WITH APPLICATIONS IN PARAMETRIC COST ESTIMATION.

(10) A.S. Korkotsides and K.T. Wallenius
Clemson University

Department of Mathematical Sciences

(7) Technical Report, #349

(11) August, 1980

(12) 19

(14) N117 TR-349

This work was supported in part by the Office of Naval Research under Contract (15) N00014-75-C-0451.

407183

K

ABSTRACT

In the context of multiple linear regression, when a subset of k-out-of-p predictor variables is to be selected for the purpose of predicting the response at some known point in the predictor variables' space, the width of the resulting prediction interval gives an indication of the precision with which the response is predicted and, thus, it may provide a suitable selection criterion.

A review of commonly used selection criteria is given, with special emphasis on those which deal with the problem of prediction. The Mahalanobis distance is one of the quantities affecting the width of the prediction interval, and it is studied in some detail. The effects of adding a new variable to a model are investigated and a monotonicity theorem is derived.

The influence of an observation on the width of the prediction interval, as measured by the effected change when that observation is set aside, is also investigated and an equivalence between observation deletions and variable augmentation is shown.

The relationships found in these investigations indicate the applicability of certain computing techniques. Computing algorithms are presented.

A management science application of the statistical procedures developed in this study is explored in the area of parametric cost estimation.

TABLE OF CONTENTS

	Page
TITLE PAGE	i
ABSTRACT	ii
DEDICATION	iv
ACKNOWLEDGMENTS	v
LIST OF TABLES	vii
LIST OF FIGURES	viii
CHAPTER	
I. INTRODUCTION	1
II. ON THE SELECTION OF VARIABLES	8
The Need for Variable Selection	8
Review of Selection Criteria	13
III. ON THE WIDTH OF THE PREDICTION INTERVAL	22
Mahalanobis Distance as a Measure of Analog	22
The Prediction Interval	31
General Observations About W	35
IV. COMPUTATION	54
A Branch and Bound Algorithm	54
A Stepwise Algorithm	57
V. ON THE INFLUENCE OF OBSERVATIONS ON W	63
Theory and Discussion	63
Computation	74
Row Deletion and Variable Augmentation - An Equivalence	77
VI. APPLICATION	80
VII. DISCUSSION AND CONCLUSIONS	104
BIBLIOGRAPHY	108

LIST OF TABLES

Table		Page
I.	Aircraft Data	82
II.	Absolute Errors in Logarithmic and Original Units	88
III.	Gain in Logarithmic and Original Units	89
IV.	Performance Statistics in Logarithmic Units ..	91
V.	Performance Statistics in Original Units	91
VI.	Average Nominal 95% Prediction Interval Widths and Frequency of Coverage in Logarithmic Units	91
VII.	Minimum W Models and Aircraft Predicted	95
VIII.	Minimum MSE Models and Aircraft Predicted	96
IX.	Minimum C_k Models and Aircraft Predicted	97
X.	Maximum F Models and Aircraft Predicted	98
XI.	R^2 Models and Aircraft Predicted	99
XII.	MSEP Models and Aircraft Predicted	100
XIII.	Observation Deleted and Maximum Reduction in Width of the Prediction Interval for Each Aircraft	102

LIST OF FIGURES

Figure		Page
1.	Extrapolation in Two Dimensions	24
2.	W-optimal Models in Two-Dimensional Space. An Illustration	38
3.	Graphic Display of Analogy Versus Fit. One-Variable Models	42
4.	Graphic Display of Analogy Versus Fit. Two-Variable Models	43
5.	Graphic Display of Analogy Versus Fit. Three-Variable Models	44
6.	Graphic Display of Analogy Versus Fit. One-Variable Models	47
7.	Graphic Display of Analogy Versus Fit. Two-Variable Models	48
8.	Graphic Display of Analogy Versus Fit. Three-Variable Models	49
9.	Graphic Display of Analogy Versus Fit. Four-Variable Models	50
10.	Graphic Display of Analogy Versus Fit. Five-Variable Models	51
11.	Graphic Display of Analogy Versus Fit. Six-Variable Models	52

CHAPTER I

INTRODUCTION

Multiple regression analysis is, probably, the most widely used and abused of all statistical tools (5). Many authors attest to the importance and wide applicability of this technique (5), (9), (28), etc. The advent of high-speed digital computers and the development of efficient software packages have made it accessible to users from all fields of research. Associated with the enhanced availability and ease of use provided by these technological developments is a tendency to apply regression techniques routinely and mechanically, without due consideration to the underlying theory or the empirical "rules of thumb" consistent with that theory and with common sense.

The main step in any regression analysis is the development of an equation relating one variable, commonly referred to as the response variable, to another set of variables, called explanatory or predictor variables. For some highly structured applications in the physical sciences, the exact form of the appropriate regression function may be known to the experimenter. In other cases, theory may specify a functional form to be tested. These cases, however, are the exception rather than the rule. More often, the analyst is uncertain about which variables are important

carriers of information, as well as about the form of the relation. In those cases, the analyst must let the data speak for itself in suggesting candidate model specifications. This process is referred to as "data mining" by Leamer (22), and is more formally known as empirical model building. Usually, at an early stage, a large number of potential predictor variables must be considered, some of which may be transformations of other variables. The task of the analyst is to bring to the surface the "nuggets of truth" which are hidden in a set of observations on the variables by means of a thorough and appropriate investigation. There are many reasons why one must be parsimonious in his use of variables. Some of them may be totally irrelevant to the problem, while others may be "conditionally irrelevant" in the sense that, in the presence of other variables they possess little or no explanatory value. It is tempting to use "all the information" of the "full" model but this often causes problems associated with what is referred to as "overfitting". Models with many variables result in large prediction variances (35) as well as statistical and computational instability in the presence of multicollinearity among the retained variables (3). Also important is the fact that a model with many variables may be difficult to interpret and/or maintain. Thus, the need arises for techniques which will screen the variables and select a subset of them deemed appropriate for the intended use of the model.

Various techniques, commonly referred to as "variable selection criteria", have been suggested for this purpose, such as minimizing the mean square error (MSE) or, equivalently, maximizing the adjusted coefficient of determination, R_a^2, maximizing F, minimizing Mallow's C_k statistic etc. In Chapter II, the need for variable selection and some of the criteria in use are discussed. All of these commonly used techniques are based on the data only through the sum of square errors (SSE). As a result, for any given number of variables, they all select the same subset, namely, the one which minimizes SSE. This, in itself, is a rather desirable property, especially when the object of the analysis is the explanation of relations among the historical data. However, as Lindley (23) emphasizes, the technique used to develop a regression equation ought to be related to the intended use. When the object of the analysis is the development of an equation which will be used in order to predict the response at a known point in the space of the predictor variables, ignoring this additional information is contrary to Lindley's recommendation and to common sense. The issue of how to use such information needs, therefore, to be investigated. The Mean Square Error of Prediction (MSEP) criterion, which is discussed in the next chapter, represents an attempt towards utilizing the information carried by the point under prediction. Its approach, however, does not seem to be fully satisfactory for several reasons

which will be discussed in subsequent chapters. Therefore, there remains a void in the literature in this respect, which this dissertation attempts to fill.

More specifically, the problem can be described as follows: A future observation on the response variable, Y, must be predicted, using the relational information provided by a set of n historical observations on Y and a set of p predictor variables $X_1, X_2, \ldots, X_p$ potentially related to it, as well as the values $\underline{x}$ of the predictor variables associated with that future observation. The relative location of $\underline{x}$ with respect to the historical data yields additional information which, if ignored, may lead to models not well suited to predict at that location.

The width of the resulting prediction interval at $\underline{x}$ is a numeraire which seems like a reasonable basis for choosing among alternative models. The Mahalanobis distance, introduced by P. C. Mahalanobis (24) as a measure of divergence between groups of multivariate data, affects the width of the prediction interval and may provide a measure of analog between $\underline{x}$ and the historical data. In Chapter III, the theoretical aspects of the problem are investigated. An interesting result which leads both to an easy geometric interpretration and to existing computing techniques is derived in the form of a monotonicity theorem. This theorem is also used in order to explain certain observations made during the simulations which were conducted and the analyses which were performed on real data sets.

The computational aspect of the problem as it relates to the proposed selection technique is investigated in Chapter IV. This is an important consideration because the need for variable selection becomes more pronounced as the number of potential predictor variables increases. An existing, efficient algorithm is modified for the purposes of this criterion, by utilizing the results of the theorem in Chapter III. Using the same theorem, stepwise FORWARD selection and BACKWARD elimination algorithms are developed.

The leverage of individual observations on the various quantities of interest should be an integral part of any analysis and has recently received deserved attention in the literature (7), (8), (13), (18), (36). Observations which seem discrepant or damaging in some sense appropriate to the analysis are allotted special attention and are investigated further. In the context of this investigation, an observation calls for such attention if its deletion from the least squares calculations results in a significant change in the width of the prediction interval. Chapter V deals with this issue. Some results are derived, some observations are made and computational formulae are given.

In Chapter VI, an application from the field of management science, referred to as parametric cost estimation, is investigated. A real data set is analyzed and the performance of this criterion is compared to that of others. The

results of a limited simulation study are also briefly discussed.

Finally, in Chapter VII, some concluding remarks are made, suggestions for the use of the new criterion are given and questions relevant to the problem at hand which were not investigated in this study are raised.

A word of warning is appropriate here, which applies to every statistical analysis of data and, in particular, to every variable selection technique. As was mentioned above, regression is one of the most widely used statistical techniques because of its wide applicability, ease of use and elegance. It is also one of the most widely abused techniques. Two of the reasons for such abuse are:

(a) The proliferation of efficient statistical packages with a variety of regression options.

(b) The lack of awareness on the part of the practitioner about the dangers of such misuse.

It may be that the practitioner has not been warned by the statistician emphatically or frequently enough. However, it remains of paramount importance that the practitioner be aware of the following:

There is no substitute for a well thought out, well executed and complete analysis. There are many sides to an analysis and data sets behave in their own peculiar ways. Standard, mechanical approaches often fail to reveal these peculiarities and, even if they do, appropriate remedial action

requires more than superficial familiarity with the model and its relation to the real world process. For example, there is no variable selection technique which is automatically applicable to all situations. Even for a given data set, there is rarely a "best" criterion or a "best" model that is known to the analyst. For a good analysis, potential variables and candidate model specifications should be decided after "eyeballing" the data, and with input from the experts in the field of application. Part of the data should be set aside for validation purposes, whenever such a luxury can be afforded. Models should be kept for further scrutiny that have good "automatic" properties such as large R^2, small C_k, small prediction intervals etc. If probabilistic statements are to be made, which is almost always the case, the residuals should be analyzed for indications of model inadequacy and of violations of the assumptions on the errors. Finally, the model (or models) passing all tests should be subjected to the criticisms of the experts in the field. In the process described above, only the computations may be done in an automatic way. The analyst's judgement and knowledge put to good use is what constitutes the difference between data analysis and the simple processing of data.

CHAPTER II

ON THE SELECTION OF VARIABLES

The Need for Variable Selection

In most practical situations, finding an equation which will describe a set of data collected in a manner referred to by Box (5) as an "unplanned experiment" is a difficult task. The problem which is investigated in this dissertation can be described as follows:

There are available n observations (fundamental measurements) on one response variable, V, denoted by V_i, $i=1,2,\ldots,n$ and n associated observations on m basic, or fundamental, variables $Z_1,Z_2,\ldots,Z_m$, denoted by $Z_{i1},Z_{i2},\ldots,Z_{im}$, $i=1,2,\ldots,n$. There is one more measurement $z_1,z_2,\ldots,z_m$ on the basic variables. An equation of the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i \qquad (2.1)$$

is assumed relating $Y_i = f(V_i)$ and $X_{ij} = g_j(Z_{i1},\ldots,Z_{im})$, $j=1,2,\ldots,p$, $i=1,2,\ldots,n$. Henceforth, the variables X_j, $j=1,2,\ldots,p$ will be referred to as explanatory, or predictor variables. This equation is assumed to be linear in the parameters $\beta_0,\beta_1,\ldots,\beta_p$ and it need not be linear in the original variables $Z_1,Z_2,\ldots,Z_m$ as g_j, $j=1,2,\ldots,p$ may be any functions of those variables. For example, $X_1=Z_1^2$,

$X_2=Z_1Z_3$, $X_3=\log Z_4$ etc., will produce an equation which is not linear in the basic variables $Z_1,Z_2,\ldots,Z_m$. In matrix terms, the model can be described by $\underline{Y} = \underline{\underline{X}}\underline{\beta} + \underline{\varepsilon}$, where $\underline{Y}$ is the $n\times 1$ vector of responses, $\underline{\underline{X}}$ is the $n\times(p+1)$ matrix of the values of the explanatory variables whose first column consists of 1's and which is assumed to be of full cclumn rank, and $\underline{\beta}$ is the p+1 dimensional vector of the unknown parameters. The object of most statistical analyses is to estimate the parameters $\underline{\beta} = [\beta_0,\beta_1,\ldots,\beta_p]'$ by means of $\underline{b} = [b_0,b_1,\ldots,b_p]'$. The estimate $\underline{b}$ is usually obtained by the method of least squares, i.e. $b_0,b_1,\ldots,b_p$ are such that

$$Y_i = b_0 + b_1X_{i1} +\ldots+ b_pX_{ip} + e_i \qquad (2.2)$$

with

$$\sum_{i=1}^{n} e_i^2 = \min_{\underline{b}}\{\sum_{i=1}^{n}[Y_i-b_0-b_1X_{i1}-,\ldots,-b_pX_{ip}]^2\}. \qquad (2.3)$$

In this investigation, the case where the resulting equation will be used to predict the response y associated with the point $\underline{x} = [x_1,\ldots,x_p]$, where $x_j=g_j(z_1,\ldots,z_m)$, $j=1,2,\ldots,p$ is considered. In these cases, the main consideration is the accurate prediction of y rather than the estimation of $\underline{\beta}$. For curve fitting purposes, the n-dimensional vectors $\underline{X}_1,\underline{X}_2,\ldots,\underline{X}_p$ are assumed independent in the algebraic sense. In order to make statistical inferences about the standard errors of the estimates, the precision of the pre-

dicted values etc., the errors ε_i are, in addition, assumed to be jointly distributed as $N(\underline{0},\sigma^2 I)$, i.e. normal, with mean vector zero and covariance matrix $\sigma^2 I$. This covariance structure implies that the errors have equal variances (homoscedastic) and are uncorrelated.

The initial choice of what aspects of the sample units ought to be measured may be straightforward, as is the case in well understood situations where physical laws and theories apply or are being tested. In less structured situations, such as behavioral research, exploratory studies often start by measuring most everything and then let the data "speak for itself" in identifying the important variables and forms of relations. Whatever process is used in assembling the set of p candidate predictors, $X_1, X_2, \ldots, X_p$, it is hoped that the list is extensive enough to include all of those which have influence on the response variable Y. To be so inclusive, the list often contains useless variables and/or variables whose informational value is superfluous in the presence of other explanatory variables. As part of the more general problem of analyzing a given set of data, subsets of variables must be selected, which seem to explain the data adequately. Selecting the essential variables is a source of trouble with unplanned data. One major reason for this is that the problem does not yield to a universal definition. What is precisely meant by saying that a model is sought which "adequately represents the data"?

Which facet of the data should the analyst ask the chosen model to represent best? The answer to such questions must depend on the intended use of the model as discussed by Lindley (23). The idea of a model which is "best" for prediction, or a model which is "best" for estimation, for instance, is elusive. Indeed, several answers to such questions might be appropriate as the problem is not one but several, intricately interwoven. It is a generally accepted maxim among statisticians, however, that parsimony in model building is desirable. There are several theoretical and practical reasons for this view as follows:

1. Models with too many variables usually result in large prediction variances due to the fact that many parameters have to be estimated. Walls and Weeks (35) have shown that the variance of prediction increases with the number of variables in the regression equation. For this reason, the analyst would like to detect and exclude those variables which are either irrelevant to the problem or unnecessary in the presence of others which are to be retained.

2. With a large number of variables, statistical instability of the resulting equation is more likely to occur. Statistical instability is the phenomenon in which a small perturbation in the values of some of the variables results in a large change in the coefficients of the fitted equation. This is one of the visible effects of multicollinearity, namely the phenomenon of strong association among the

retained variables. Mathematically, this phenomenon occurs when the matrix $\underline{X}'\underline{X}$ is nearly degenerate. The phenomenon of multicollinearity can appear because the data come from a subspace of the true sample space, one that can almost be described in fewer dimensions. This may happen either by chance, or by necessity, or by the inclusion of extraneous variables which are strongly associated with the relevant predictors (21), (27). In such cases, the estimates of the regression coefficients have large variances resulting in instability of the hyperplane defined by the regression equation. This is easy to visualize in the case of two highly correlated explanatory variables. If the data are nearly collinear (one-dimensional subspace), the regression plane is "resting on a knife's edge". Any perturbation in the data can make it tilt heavily. Again, it might be desirable to use a subset of variables so as to alleviate the problem, especially if the variables which are causing the multicollinearity are extraneous anyway.

3. Another undesirable effect of multicollinearity is computational instability, resulting in potentially large roundoff errors (3).

4. Finally, from the purely practical point of view, a model with many variables may be difficult to interpret, difficult or costly to maintain, or both. Interpretation of relations between individual predictor variables or groups of them and the response variable is often desirable, and

collection of data on certain variables is often difficult, unreliable or costly.

The reasons mentioned above should suffice in explaining why the problem of variable selection is real and, often, of great practical importance. In the next section, some commonly applied selection techniques are discussed.

Review of Selection Criteria

For this section and the ones that follow, some new notation will be needed. The i-th response is estimated by

$$\hat{Y}_i = b_0 + b_1 X_{i1} + \ldots + b_p X_{ip} \quad i=1,2,\ldots,n. \qquad (2.4)$$

The i-th residual is defined by $e_i = Y_i - \hat{Y}_i$ and the sum of squared errors (residual sum of squares) is defined by

$$SSE = \sum_{i=1}^{n} e_i^2 . \qquad (2.5)$$

In variable selection the possibility of setting some of the p coefficients equal to zero is considered. This amounts to selecting a subset of k, say, out of the p variables. The mean of the n observations on the response variable is denoted by $\bar{Y}$, the total sum of squares of deviations from that mean is defined by

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \qquad (2.6)$$

and the regression sum of squares by

$$SSR = SSTO - SSE \qquad (2.7)$$

A selection criterion is a rule according to which a certain model out of the 2^p possible models is labeled "best". It should be noted that "best" is defined only in the sense of the particular criterion employed, and it does not necessarily imply that that model is best in terms of its intended use or in terms of how well the relation defined by it carries over to the population. The position taken in this dissertation concerning variable selection and model building is more general, namely, that "selection" rules ought to be used in order to screen the 2^p models down to a more manageable number, say half a dozen or so, which, subsequently, would be carefully scrutinized for adequacy and reasonableness. There are several criteria currently used for this purpose. The most common ones, as well as some which are related to the problem of prediction are discussed next.

1. The R^2 Criterion

The coefficient of determination is defined by

$$R^2 = 1 - SSE/SSTO. \tag{2.8}$$

It is clear that R^2 is the proportion of variability in Y which is explained by the variables in the model under consideration. It seems desirable that, other things being equal, R^2 should be as large as possible. However, since SSE can not increase as variables are introduced into the model, R^2 will always achieve its maximum when all p

variables are used. If R^2 is to be used as a selection criterion, some subjective rule must be employed that will determine when the largest increase possible in R^2 attained by the introduction of a new variable does not compensate for the loss in degrees of freedom due to estimating an additional parameter. A graph of R^2 versus model size is usually helpful in devising what is called an "elbow rule".

2. The Adjusted R^2 Criterion (Mean Square Error)

To overcome the subjectivity involved in using R^2, an adjustment for degrees of freedom can be made by defining the adjusted coefficient of determination, R_a^2, by

$$R_a^2 = 1 - [SSE/(n-k-1)]/[SSTO/(n-1)] \qquad (2.9)$$

where k is the number of predictor variables in the postulated model. This statistic usually achieves a maximum with a model containing fewer than p variables. The equation

$$R_a^2 = R^2 - k(1-R^2)/(n-k-1) \qquad (2.10)$$

shows the relationship between the statistics R^2 and R_a^2. This criterion is equivalent to selecting the model with smallest mean square error, defined by MSE = SSE/(n-k-1), since the denominator in R_a^2 does not change with the variables selected. A preference for choosing models with large R_a^2 is based on the fact that the "true" model minimizes the expected MSE (32). This criterion is most often referred to

as the "minimum Mean Square Error" criterion, and this name will be used in what follows.

3. The Maximum F Criterion

Sometimes it is deemed desirable to maximize the ratio

$$F = [(SSTO-SSE)/k]/[SSE/(n-k-1)]. \qquad (2.11)$$

The numerator in the expression above is referred to as the regression mean square. This criterion is used less frequently than the others, and it is very parsimonious. That is to say, it tends to select models with very few variables.

4. Mallow's C_k Criterion

Mallows (25), introduced the statistic

$$C_k = SSE/\hat{\sigma}^2 + (2k-n) \qquad (2.12)$$

where $\hat{\sigma}^2$ is an estimate of σ^2. C_k is an estimate of the standardized total squared error of predicting at the points in the data base (19). A model with small bias is expected to yield a C_k statistic about equal to the number of variables, k, associated with it as can easily be shown. In this investigation, the model with smallest C_k will be referred to as the "minimum C_k" model and will be used for comparison purposes. Usually, the MSE of the model containing all variables under consideration is used for $\hat{\sigma}^2$, although this forces C_p to be equal to p. Easily interpretable plots of

C_k versus k can be drawn and it is suggested that models with small C_k be considered. The C_k statistic and its properties have been discussed by Daniel and Wood (9), Gorman and Toman (15), Hocking (19), Mallows (25), (26) and others.

All of the criteria discussed above share two properties.

(a) They are all simple functions of SSE and, thus, for any fixed number of variables, they all select the same model, namely the one which minimizes SSE.

(b) If the final model is to be used in order to predict the response y at a known point $\underline{x}$ in the space of the predictor variables, they all ignore its location and its characteristics with respect to the historical data. As Wallenius (34) pointed out, "...the first of these properties is reasonable but myopic when the object is prediction. The second one seems contrary to all reason."

Of the four criteria, Mallow's C_k technique is more directly related to the problem of prediction in view of the fact that it utilizes the total square error of prediction.

5. The Prediction Sum of Squares Criterion

David M. Allen (2) suggested the following selection procedure:

Let $\hat{Y}_i^{(i)}$, i=1,2,...,n denote the i-th predicted response, when a given model is used, and with the i-th observation

removed from the data base, so that the coefficients are derived from the least squares calculations based only on the remaining n-1 observations. For each model, compute the prediction sum of squares (PRESS) statistic, given by

$$\text{PRESS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i^{(i)})^2. \qquad (2.13)$$

Consider models with small PRESS. Notice that PRESS is an indication of the predictive performance of a model over the points in the data base. Intuitively, a model with small PRESS should be expected to do better in predicting future observations than a model with large PRESS. However, since this technique fails to take into account the values of the variables at the point under prediction, it is conceivable that there can be points, both in the region of the historical data and outside, where the selected model may be inappropriate.

In terms of the computational aspect of the problem, this method is much more demanding than the four previously mentioned.

6. Mean Square Error of Prediction

The mean square error of prediction (MSEP) of the response y at a given point $\underline{x}$ can be expressed as

$$E(y-\hat{y})^2 = \sigma^2 + \text{Var}(\hat{y}) + (\text{bias})^2, \quad \text{i.e.} \qquad (2.14)$$

$$E(y-\hat{y})^2 = \sigma^2 + \underline{x}(\underline{X}'\underline{X})^{-1}\underline{x}'\sigma^2 + [E(y)-E(\hat{y})]^2. \qquad (2.15)$$

Allen (1) proposed choosing subsets of variables which minimize an estimate of the mean square error of prediction. This is a difficult task to accomplish successfully, mainly because of the fact that a good estimate of the bias term assumes knowledge of E(y) or, at least, a very good estimate of it. However, this is at the very heart of the problem of prediction which the analysis attempts to solve. An assumption about good knowledge of E(y) seems to create a logical vicious circle in that the unknown answer to the problem is somehow used in order to get to it. Allen's approach to this is to assume that the full model contains variables which were chosen carefully, so as to include all relevant ones and exclude all unnecessary ones. As a result of this, the full model will be unbiased, while any submodel will be biased to a measurable degree. The bias associated with a given submodel is then estimated by the difference in point predictions between the full model and the submodel, an estimate of σ^2 based on the sum of squared errors of the full model is used in the expression for $E(y-\hat{y})^2$ and the submodel is found preferable to the full model if and only if the reduction in prediction variance is greater than the square of the bias.

Even with the assumptions mentioned above, the method of estimating the bias (a difference in expectations) by means of a difference in two point predictions may result in treating different submodels unfairly. The degree of this

unfairness will depend on the difference between E(y) and the point prediction obtained from the full model. It should also be observed that the MSE's of the postulated submodels are not taken into consideration. As a result, the selected model may provide a very poor fit to the data. As will be discussed in Chapters III and VI, this seems to be frequently the case in practice. With regard to the computational aspect, Allen proposed a sequential procedure which provides no guarantee that an absolute minimum will be obtained, either overall or for a fixed subset size k. It seems that, for such a guarantee, a complete search of all 2^p-1 regressions might be necessary. However, the algorithm which will be developed in Chapter IV may be modified so as to apply to the MSEP criterion .

In the next chapter, a somewhat different approach is taken to the problem described above. The position taken in this dissertation with respect to bias is the following: When the true population model is known, the bias at $\underline{x}$ associated with other models can be obtained. In empirical work, however, where the true model is not only unknown but its notion is not even easily or well defined, that bias can not be objectively measured. Indirect methods may be employed that can, hopefully, give indications of its magnitude. It is believed, however, that using such estimates directly in the screening of variables is risky at best and rather inapropriate. Thus, no explicit attempt is made to

estimate bias during the selection. Implicitly, bias is hoped to be reflected in the size of MSE which will be used as an estimate of σ^2.

CHAPTER III

ON THE WIDTH OF THE PREDICTION INTERVAL

Mahalanobis Distance as a Measure of Analog

The Mahalanobis distance, introduced by P. C. Mahalanobis (24) as a measure of the distance between two multivariate populations, is a fundamental notion in multivariate statistics. In this section, the Mahalanobis distance is discussed in relation to the problems of prediction and variable selection. The insight gained through this investigation will be used to explain certain observations which were made during the course of this study and to derive computational algorithms for implementation of the methodology developed.

Suppose there are n historical observations on p potential predictor variables, $X_{i1}, X_{i2}, \ldots, X_{ip}$, $i=1,2,\ldots,n$. Let $\underline{Y}$ denote the $n\times 1$ vector consisting of the observed values of the response variable, associated with these n observations, and let $\underline{X}_j$ denote the $n\times 1$ vector of observed values on variable X_j, i.e., $\underline{Y} = [Y_1, Y_2, \ldots, Y_n]'$ and $\underline{X}_j = [X_{1j}, X_{2j}, \ldots, X_{nj}]'$. Let $\underline{\underline{X}}$ denote the $n\times p$ matrix whose columns are $\underline{X}_j$, $j = 1,\ldots,p$, and let $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ be the sample mean of variable X_j. The point $\underline{\bar{X}} = [\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p]$ is defined as the centroid of the data. Finally, let

$$S = \{s_{ij}\}, \quad i,j=1,\ldots,p$$

denote the sample covariance matrix of variables $X_1,\ldots,X_p$

$$(\text{i.e., } s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} [X_{ki}-\bar{X}_i][X_{kj}-\bar{X}_j]).$$

Denoting the values of the point under prediction by lower case letters, the response y at the point $\underline{x}$ must be predicted by exploiting the predictive relationship between Y and the characteristics $X_1,\ldots,X_p$, and the degree of analogy between $\underline{x}$ and $\underline{X}$. In general geometric terms, "degree of analogy" refers to the position of $\underline{x}$ relative to $\underline{X}$ in p-dimensional space. If $\underline{x}$ is far removed from the historical data, extrapolation is necessary with all the attendant risks. This point will be discussed in more detail in the next section. The issue of how to detect such extrapolation is considered first.

The standard Euclidean distance may fail to reveal the degree of extrapolation, due to the intercorrelations among the variables. Points at a small Euclidean distance from the centroid $\bar{\underline{X}}$ of the data may be very non-analogous in that their coordinates do not conform with the correlation structure observed in the historical data. This point can be illustrated in two dimensions. In Figure 1 below, the point $\underline{x}$ is at a rather small Euclidean distance from the centroid $\bar{\underline{X}}$. Nevertheless, it is well outside the bulk of the data, because its coordinates do not conform with the negative correlation observed.

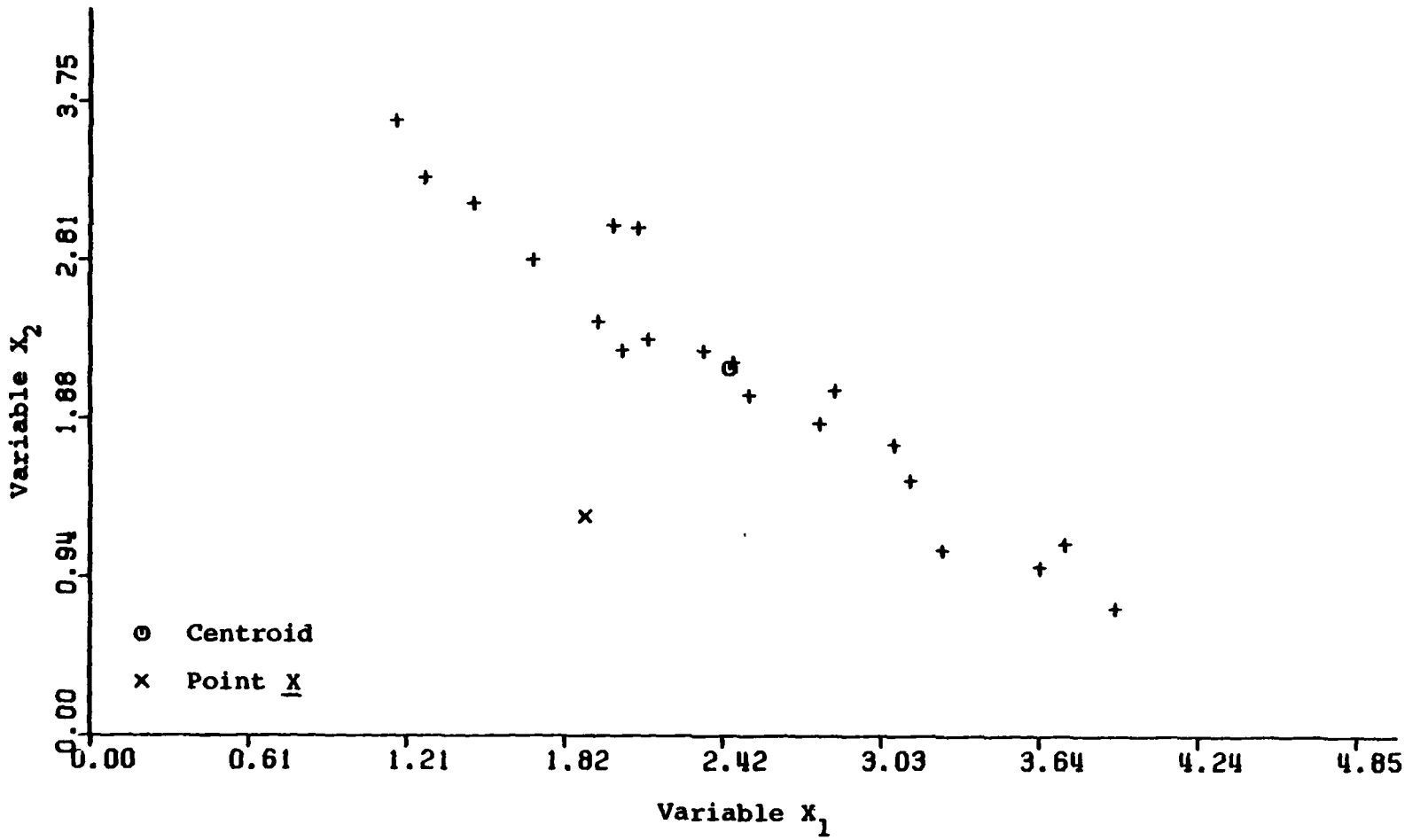


Figure 1. Extrapolation in Two Dimensions.

Observe also that the extrapolation in the two-dimensional scatter would not have been revealed by simple marginal comparisons. Each coordinate of $\underline{x}$ is well inside the range of the data along the corresponding dimension. Of course, in two or three dimensions, scatter diagrams can be plotted, which will reveal this phenomenon. In higher dimensions, a different method becomes necessary.

The Mahalanobis distance defined by

$$D(\underline{\mu}_1,\underline{\mu}_2,\Sigma) = (\underline{\mu}_1-\underline{\mu}_2)\ \Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)' \qquad (3.1)$$

is a measure of the distance between two multivariate populations with row vector means $\underline{\mu}_1$ and $\underline{\mu}_2$ and common positive definite covariance matrix Σ. The degree of analogy (distance) between $\underline{x}$ and the data $\underline{\underline{X}}$ can be described by means of the sample counterpart of the above measure, namely

$$M = (\underline{x}-\bar{\underline{X}})S^{-1}(\underline{x}-\bar{\underline{X}})'. \qquad (3.2)$$

The measure M will be referred to as the Mahalanobis distance between $\underline{x}$ and $\bar{\underline{X}}$. Observe that, except for a multiplicative constant, this is Hotelling's T^2 statistic used to test the hypothesis that $\underline{x}$ and the historical data come from the same multinormal population, assuming equal covariances. In the univariate case, M is (a multiple of) a squared t-ratio. In the multivariate case as well, M can be viewed as the square of a t-ratio. It is the squared t-ratio of that linear combination of the variables which produces

the largest t-ratio. Each univariate t-ratio corresponds to one such linear combination. However, as mentioned earlier, marginal, univariate comparisons may fail to reveal the degree of extrapolation. All univariate t-ratios may be small, although the multivariate Mahalanobis distance may be arbitrarily large. For instance, with only two variables, M can be expressed as

$$M = \frac{M_1 - 2r(M_1 M_2)^{1/2} + M_2}{1-r^2} \qquad (3.3)$$

where M_1 and M_2 are the corresponding univariate measures, and r is the sample correlation coefficient between the two variables. It is clear that, even if both M_1 and M_2 are small, M can still be large. For example, if $M_1 = M_2 = \varepsilon$, then $M = 2\varepsilon/(1+r)$, which can become arbitrarily large with $r \to -1$.

A few things might be of interest to note about the Mahalanobis distance. Points equidistant from the centroid $\underline{\bar{X}}$ form ellipsoids with center at $\underline{\bar{X}}$ whose axes coincide with the principal components axes of the data. It is a distance measure, that is, it is non-negative, symmetric, and satisfies the triangular inequality. If S = I, the Mahalanobis distance becomes the natural Euclidean distance in p-space. For an arbitrary positive definite S, the Mahalanobis distance is equivalent to the Euclidean distance in the "Mahalanobis space" $S^{1/2}\underline{x}'$, where $S^{1/2}$ is the symmetric square root of S.

With respect to this investigation, the behavior of the Mahalanobis distance as variables enter or leave the regression equation is of interest. The monotonicity theorem which follows allows the expression of the change in the Mahalanobis distance as a new variable is introduced, in terms of easily recognizable regression statistics and provides insight which will aid in subsequent analysis.

Theorem 3.1: Let M_k denote the Mahalanobis distance between $\underline{x}_k = [x_1,\ldots,x_k]$ and $\underline{\bar{X}}_k = [\bar{X}_1,\ldots,\bar{X}_k]$. Let S_{11} denote the covariance matrix of variables $X_1,\ldots,X_k$. Let M_{k+1} denote the Mahalanobis distance between $\underline{x}_{k+1} = [\underline{x}_k, x_{k+1}]$ and $\underline{\bar{X}}_{k+1} = [\underline{\bar{X}}_k, \bar{X}_{k+1}]$ and let $\Delta M = M_{k+1} - M_k$. Partition the covariance matrix of $X_1,\ldots,X_k,X_{k+1}$ as

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} .$$

Then,

$$\Delta M = \frac{(x_{k+1} - \hat{x}_{k+1})^2}{S_{22}(1-r^2)} \qquad (3.4)$$

where

$$\hat{x}_{k+1} = \bar{X}_{k+1} + S_{21} S_{11}^{-1} (\underline{x}_k - \underline{\bar{X}}_k)'$$

and r is the multiple correlation coefficient between variables X_{k+1} and $X_1,\ldots,X_k$.

Proof: From a well known matrix identity (16),

$$S^{-1} = \begin{bmatrix} [S_{11}-S_{12}S_{22}^{-1}S_{21}]^{-1} & -S_{11}^{-1}S_{12}[S_{22}-S_{21}S_{11}^{-1}S_{12}]^{-1} \\ -S_{22}^{-1}S_{21}[S_{11}-S_{12}S_{22}^{-1}S_{21}]^{-1} & [S_{22}-S_{21}S_{11}^{-1}S_{12}]^{-1} \end{bmatrix}$$

and

$$[-S_{22}^{-1}S_{21}(S_{11}-S_{12}S_{22}^{-1}S_{21})^{-1}]' = -S_{11}^{-1}S_{12}[S_{22}-S_{21}S_{11}^{-1}S_{12}]^{-1}.$$

Therefore,

$$\begin{aligned} \Delta M = {} & (\underline{x}_k-\underline{\bar{X}}_k)[S_{11}-S_{12}S_{22}^{-1}S_{21}]^{-1}(\underline{x}_k-\underline{\bar{X}}_k)' - \\ & (x_{k+1}-\bar{X}_{k+1})S_{22}^{-1}S_{21}[S_{11}-S_{12}S_{22}^{-1}S_{21}]^{-1}(\underline{x}_k-\underline{\bar{X}}_k)' - \\ & (\underline{x}_k-\underline{\bar{X}}_k)S_{11}^{-1}S_{12}[S_{22}-S_{21}S_{11}^{-1}S_{12}]^{-1}(x_{k+1}-\bar{X}_{k+1}) + \\ & (x_{k+1}-\bar{X}_{k+1})[S_{22}-S_{21}S_{11}^{-1}S_{12}]^{-1}(x_{k+1}-\bar{X}_{k+1}) - \\ & (\underline{x}_k-\underline{\bar{X}}_k)S_{11}^{-1}(\underline{x}_k-\underline{\bar{X}}_k)' \\ = {} & (\underline{x}_k-\underline{\bar{X}}_k)[S_{11}-S_{12}S_{22}^{-1}S_{21}]^{-1}(\underline{x}_k-\underline{\bar{X}}_k)' - \\ & 2(x_{k+1}-\bar{X}_{k+1})[S_{22}-S_{21}S_{11}^{-1}S_{12}]^{-1}S_{21}S_{11}^{-1}(\underline{x}_k-\underline{\bar{X}}_k)' + \\ & (x_{k+1}-\bar{X}_{k+1})^2[S_{22}-S_{21}S_{11}^{-1}S_{12}]^{-1}-(\underline{x}_k-\underline{\bar{X}}_k)S_{11}^{-1}(\underline{x}_k-\underline{\bar{X}}_k)'. \end{aligned} \tag{3.5}$$

Notice that the correlation coefficient, r, can be written as

$$r = (S_{21}S_{11}^{-1}S_{12}/S_{22})^{1/2}, \text{ with } 0 \leq r^2 \leq 1.$$

Therefore, relation (3.5) above becomes

$$\Delta M = (\underline{x}_k-\underline{\bar{X}}_k)[S_{11}-S_{12}S_{22}^{-1}S_{21}]^{-1}(\underline{x}_k-\underline{\bar{X}}_k)' -$$
$$2(x_{k+1}-\bar{X}_{k+1})[S_{22}(1-r^2)]^{-1}S_{21}S_{11}^{-1}(\underline{x}_k-\underline{\bar{X}}_k)' +$$
$$(X_{k+1}-\bar{X}_{k+1})^2[S_{22}(1-r^2)]^{-1}-(\underline{x}_k-\underline{\bar{X}}_k)S_{11}^{-1}(\underline{x}_k-\underline{\bar{X}}_k)'. \tag{3.6}$$

Using the fact (see (30)) that

$$[A+UV']^{-1} = \frac{A^{-1}-A^{-1}UV'A^{-1}}{1+V'A^{-1}U} \tag{3.7}$$

with

$$U = -S_{12}S_{22}^{-1}, \quad V = S_{12} \quad \text{and} \quad A = S_{11},$$

it follows that

$$[S_{11}-S_{12}S_{22}^{-1}S_{21}]^{-1} = S_{11}^{-1} + \frac{S_{11}^{-1}S_{12}S_{21}S_{11}^{-1}}{S_{22}-S_{21}S_{11}^{-1}S_{12}}. \tag{3.8}$$

Thus, relation (3.6) becomes

$$\Delta M = (\underline{x}_k-\underline{\bar{X}}_k)(S_{11}^{-1}S_{12}S_{21}S_{11}^{-1})[S_{22}(1-r^2)]^{-1}(\underline{x}_k-\underline{\bar{X}}_k)' -$$
$$2(x_{k+1}-\bar{X}_{k+1})[S_{22}(1-r^2)]^{-1}S_{21}S_{11}^{-1}(\underline{x}_k-\underline{\bar{X}}_k)' +$$
$$(x_{k+1}-\bar{X}_{k+1})^2[S_{22}(1-r^2)]^{-1} -$$
$$(\underline{x}_k-\underline{\bar{X}}_k)S_{11}^{-1}(\underline{x}_k-\underline{\bar{X}}_k)' + (\underline{x}_k-\underline{\bar{X}}_k)S_{11}^{-1}(\underline{x}_k-\underline{\bar{X}}_k)'$$
$$= \frac{\{x_{k+1}-[\bar{X}_{k+1}+S_{21}S_{11}^{-1}(\underline{x}_k-\underline{\bar{X}}_k)']\}^2}{S_{22}(1-r^2)} \tag{3.9}$$

QED.

Observe, first, that $\Delta M \geq 0$. Thus, as variables are introduced into the regression, the Mahalanobis distance cannot decrease. The resulting increase is equal to the standardized square error of predicting x_{k+1} from the regression of the newly introduced variable X_{k+1} on the variables $X_1, X_2, \ldots, X_k$ which were already in the regression equation. The standardization is done by dividing the resulting squared error by the conditional variance of variable X_{k+1} (conditioned on variables $X_1, X_2, \ldots, X_k$). This standardization implies that the expected change in M should not depend on the strength of the relationship between X_{k+1} and $X_1, X_2, \ldots, X_k$. Obviously, $\Delta M = 0$ if and only if x_{k+1} is on the hyperplane defined by the regression mentioned above. The Mahalanobis distance is a unitless quantity whose size does not depend on the units of the particular problem. For a given problem, $\underline{x}$ and $\underline{\underline{X}}$ are, of course, fixed. Over all $\underline{x}$ and $\underline{\underline{X}}$, however, drawn from a k-variate normal population, the quantity:

$$\frac{n}{n+1} + \left[\frac{1}{n-1}(\underline{x}_k - \bar{\underline{X}}_k) S^{-1} (\underline{x}_k - \bar{\underline{X}}_k)'\right]$$

is distributed like Hotelling's two-sample T^2, and, so, it has the distribution of

$$\frac{k(n-1)}{n-k} F_{k,n-k'}$$

where $F_{k,n-k}$ is an F variate with k and n-k degrees of freedom. This distributional property of M implies that its expected magnitude depends only on the number of

variables, k, and is independent of the specific variables involved. Thus, the relative magnitude of the realized M for a particular subset of variables can be assessed.

The result of this theorem will be used in what follows and, in particular, in Chapter IV where the computational aspect of the problem is investigated.

The Prediction Interval

As was mentioned earlier, all standard variable selection techniques share the property that, for any given number of variables in the regression equation, the optimal set is the one which minimizes the sum of squared residuals or, equivalently, maximizes R^2. The point under prediction may be rather non-analogous to the historical data (large M) when we consider the set of variables identified as "optimal" by the criterion used. In such cases, the model will be required to extrapolate. The term "extrapolation" is used here in the sense that the variance of prediction

$$\sigma^2 \underline{x}(\underline{X}'\underline{X})^{-1}\underline{x}' = \sigma^2[\frac{1}{n} + \frac{M}{n-1}] \qquad (3.10)$$

is large, relative to the inherent error variability σ^2. Extrapolation should be avoided whenever possible for two reasons:

1. The hyperplane defined by the regression equation may fit the available data rather well, but this may be true only in the region of the X-space in which data are available. The true model for the full X-space may well

be quite different in the vicinity of $\underline{x}$ thus producing substantial bias.

2. Even if the variables used are the ones generating the response values Y, the variance of prediction and, as a consequence, the errors at points removed from the bulk of the data may be large due to the variances associated with the estimates $b_0, b_1, \ldots, b_k$. These variances may be large, compared with σ^2, especially in the presence of multicollinearity among the retained variables.

Ideally, variables which are extraneous to the problem as well as variables whose presence does *not* contribute significantly to the explanation of the variability in Y should be detected. Dropping such variables from the regression equation has the effect of reducing the variance of prediction at $\underline{x}$. This, of course, should not be done at the expense of excluding variables whose inclusion would greatly enhance the fit of the data as measured by the mean square error. Often, there are several models which come close to the "optimum" in terms of R^2 and other measures of model aptness based on residual analysis. In those cases, by using a slightly sub-optimal set of predictor variables (slight decrease in R^2), it may be possible to substantially improve the degree of analogy (decrease M) and thus reduce prediction variance.

To illustrate this point, suppose that two single-variable models are to be compared in terms of their expected predictive performance at $\underline{x} = [10,0]$. Suppose,

moreover, that the following statistics are associated with each model and the corresponding variables:

$$R^2_{Y|X_1} = 0.90, \quad \bar{X}_1 = 10, \quad S_{X_1} = 4,$$

$$R^2_{Y|X_2} = 0.92, \quad \bar{X}_2 = 10, \quad S_{X_2} = 4.$$

Suppose further that there are n = 10 observations on X_1, X_2 and Y, and that $S_Y = 4$. If the corresponding mean square errors, MSE_i, are used to estimate σ^2, the prediction variances

$$MSE_i x_i (X_i' X_i)^{-1} x_i', \quad i = 1,2$$

are equal to 0.180 for the first model and 1.144 for the second one. Notice that the corresponding Mahalanobis distances are $M_1 = 0$ and $M_2 = 6.25$. Even though the second variable results in a slightly better fit for the data, the point $\underline{x} = [10,0]$ is so non-analogous on this dimension that it might be preferable to use variable X_1 for prediction.

The width of the $100(1-\alpha)\%$ prediction interval at $\underline{x}$ is a numeraire which reflects the situation discussed above and, thus, it may provide a reasonable basis for choosing among competing models. For computational simplicity, a monotone function of the width will be used, namely the square of the half-width, viz.

$$W = F_{1-\alpha;1,n-k-1} MSE \left[\frac{n+1}{n} + \frac{M}{n-1}\right] \qquad (3.11)$$

where $F_{1-\alpha;1,n-k-1}$ is the α-th fractile of an F distribution with 1 and n-k-1 degrees of freedom. This measure, W, combines fit (MSE) and degree of analogy (M), with a factor F which penalizes for using too many variables (increasing k) or excluding points from the data base (decreasing n). In this form, the role of analogy as measured by the Mahalanobis distance becomes evident. Failure to consider this factor in selecting a set of predictor variables could have a marked effect on predictor precision as measured by the width of the prediction interval, and, as a consequence, on the accuracy of prediction as measured by the prediction error.

As mentioned in the previous chapter, the position taken in this dissertation is that bias is not an issue that can be dealt with directly in unstructured situations with which empirical model building is concerned, since the population model (true model) is unknown. Nevertheless, it might be of interest to note that the mean square error of prediction of the elusive population model is

$$E(y-\hat{y})^2 = \sigma^2_{Y|X}[\frac{n+1}{n} + \frac{M}{n-1}]. \qquad (3.12)$$

Thus, if one were willing to assume that the postulated empirical model is the same as the population model, then the last two factors in W would be an estimate of the mean square error of prediction. Notice that the factor $F_{1-\alpha;1;n-k-1}$ is only a penalizing factor for lost error

degrees of freedom. For any given set of n observations and any number of variables, k, the set of predictor variables which minimizes W, minimizes also this estimate of $E(y-\hat{y})^2$.

General Observations About W

The quantity W has been defined as "the squared half-width of a $100(1-\alpha)\%$ prediction interval at $\underline{x}$". Its statistical validity as a bona fide $100(1-\alpha)\%$ prediction interval is vitiated if the model is selected by minimizing W, just as the distributional properties of R^2 are no longer valid when the data is used to build a model which minimizes MSE (11). After the data have been looked into and, say, the model with smallest W is selected, the confidence associated with that interval will be less than $100(1-\alpha)\%$. Therefore, an interval centered at y with width $2W^{1/2}$ should not be thought of as a $100(1-\alpha)\%$ prediction interval but only as a relative indication of the predictive performance of the various models. Although no concrete statements can be made, it is hoped that the improvement in precision will be accompanied by an improvement in prediction accuracy. For this reason, the quantity will be referred to as "W" instead of "prediction interval" in what follows.

Another important point is that, even when the model is specified in advance, the validity of the formula used for the prediction interval rests on the usual assumptions on the errors as they were stated in the introduction. If the residuals associated with a particular model indicate

gross violations of those assumptions on the part of the errors, W becomes a meaningless statistic. Judging the predictive performance of various models on the basis of such a statistic would be quite speculative at best. Therefore, when W is employed in order to select subsets of variables, it is important that a check on the assumptions be made. Appropriate transformations on the variables should be made before judgement on the basis of W is attempted. This observation is supported by the analyses on data sets which will be discussed in Chapter VI.

Given that M cannot decrease as variables are added to a model, W may decrease only if the mean square error decreases by an amount sufficient to offset the increase in M and F. Thus, augmenting a k-variable model to reduce W will always reduce MSE, so that W-optimal models will tend to contain fewer variables than MSE-optimal models. This, in itself, is a rather desirable property in view of the commonly held opinion among statisticians that the minimum MSE criterion frequently results in considerable overfitting. Of course, this parsimony of the "minimum W" criterion is not guaranteed. The selection may take different paths for the two criteria. The opposite phenomenon was observed in only two occasions in the data which were analyzed.

For large n, W will be dominated by the factor MSE, since M is divided by n-1. This agrees with our intuition since, for a given M, the point $\underline{x}$ will be inside the

k-dimensional scatter of the data if n is large more so than if n is small. Thus, a hyperplane which explains the data well should be expected to predict the new point well too. Otherwise, for a given MSE and a given M, extrapolation as it was defined earlier, is more extreme with a small n than with a large one. The increased influence of M as n decreases, however, may have an adverse effect on the fit. A variable may be excluded that is found desirable on other considerations. It may be prudent in such cases to consider a slightly W-suboptimal model by forcing the desirable variable into the regression equation. Investigations confirm that such an occurrence is possible. At the same time, it was found that a careful analysis will reveal these anomalies. An investigation into the variables which are excluded by the minimum W criterion may provide insight into aspects of the problem which, otherwise, would not have been gained.

The W criterion partitions the p-dimensional X-space into well defined and clearly bounded regions in which different models are optimal. For purposes of illustration, a simple two varible example was used in order to obtain insight into the nature of the various regions. The result is depicted in Figure 2. Six observations on two predictors X_1 and X_2 and the response variable were used, marked by "+". For each one of the four models containing the constant term, W was expressed as a function of X_1 and X_2. The "equi-W" curves (level curves) were computed and drawn.

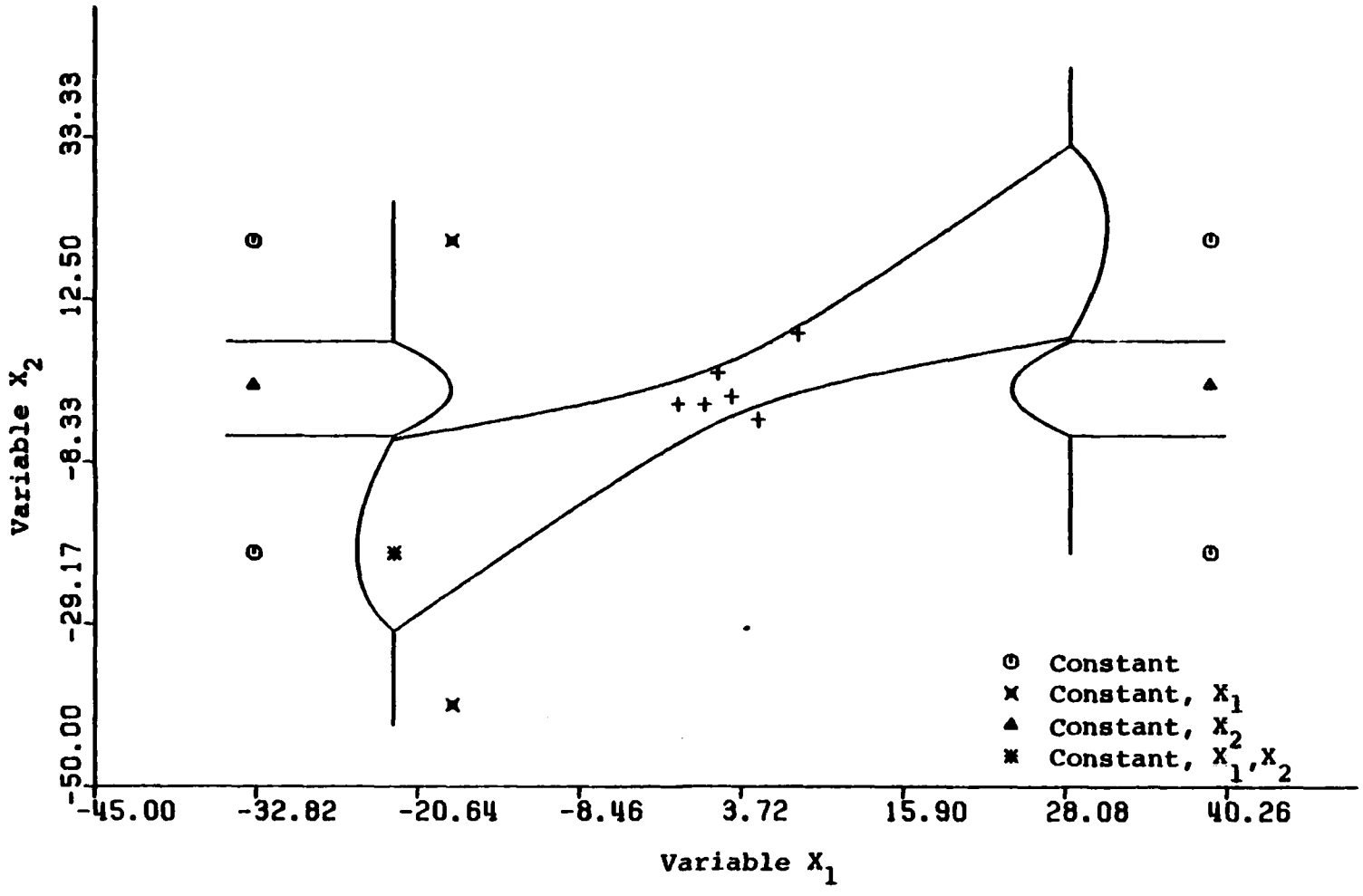


Figure 2. W-optimal Models in Two-Dimensional Space. An Illustration.

The model producing the smallest W in each of the regions defined by the level curves was found.

Notice that, as $\underline{x}$ traverses any of the boundaries, the model selected by W changes. As a result, point predictions change in a discontinuous way as a boundary is crossed. This is a somewhat disconcerting property of the criterion, even though W changes continuously. As was emphasized earlier, however, W is offered as a screening aid and not as a method for determining one, and only one, model to be labeled "best". Seen in this light, for certain points $\underline{x}$, the existence of more than one model with almost equal W's should indicate the need for further investigation of their properties.

The interpretation of the situation for points in the regions where no variable is retained also merits attention. A point $\underline{x}$ in such a region is very non-analogous to the historical data (large M). Yet, if only the constant is used, as is suggested by W, the point prediction will be none other than the mean of the historical response values. The analyst should view this occurrence as a suggestion that $\underline{x}$ and $\underline{X}$ are sufficiently non-analogous so as to vitiate the entire regression approach to prediction in the situation at hand, at least if the regression is to be based on the given body of data and the predictor variables under consideration. Even though one can obtain a good fit between Y and $\underline{X}$ in the historical data, there is no strong justification in expecting y to be analogous to Y if $\underline{x}$ and

$\underset{\sim}{X}$ were generated by different processes. Thus, a phenomenon which seemed anomalous at first glance, acts as a valuable warning for the analyst using the W criterion. In fairness to more standard approaches to prediction it is acknowledged that the careful analyst should become aware that something is amiss upon observing the large prediction interval at $\underset{\sim}{x}$ based on his selected "best fitting" model.

As mentioned earlier, W combines fit with analogousness. It is often desirable to know the relative sizes of these two factors for a given model. A graphic display can yield insight into data and enable the analyst to perceive patterns in them which might be difficult to perceive from numerical procedures and tabular displays. In the situation at hand, such a display of the magnitudes of M, MSE and W could help the analyst choose from among several competing models, according to his judgement of the relative importance of each factor. For each subset size k such a display can be constructed by first observing that MSE and W are expressed in squared Y units. In order to get unitless quantities, note that

$$MSE = S_Y^2(n-1)(1-R^2)/(n-k-1) \tag{3.13}$$

and, therefore,

$$\frac{W(n-k-1)}{F_{1-\alpha;1,n-k-1}S_Y^2} = (1-R^2)\left[M + \frac{n^2-1}{n}\right] . \tag{3.14}$$

Taking natural logarithms of both sides of (3.14) and rearranging terms,

$$\ell n[M + \frac{n^2-1}{n}] = \ell n[\frac{W(n-k-1)}{F_{1-\alpha;1,n-k-1}S_Y^2}] - \ell n(1-R^2). \quad (3.15)$$

So, for fixed k, points representing models with equal W's lie on a line with slope -1, on a graph of $\ell n[M + (n^2-1)/n]$ versus $\ell n(1-R^2)$. The intercept of such lines is determined by W. For a given W, models with small MSE and large M will be located high on the "equi-W" line, while models with large MSE and small M will be located on its lower part. For a clearer picture of the relative sizes of W across model sizes, k, these lines may be labeled by W (or the width of the prediction interval). Since $\ell n(1-R^2) \leq 0$, it might be preferable to set the origin at, say, (-5,0), so as to have most of the points in the first quadrant. Two examples were used, differing on the total number of variables involved. For the first example, the data on page 366 in Draper and Smith (10) were used. This data set involves thirteen observations on four predictor variables. The response variable measure the heat evolved during the hardening of cement containing chemical substances which are measured by the four predictors. The last row was deleted from the data base and predicted. The resulting plots for 1, 2 and 3-variable models are shown in Figures 3, 4 and 5 respectively, with the lines labeled by the width of the prediction interval. Models which were not selected

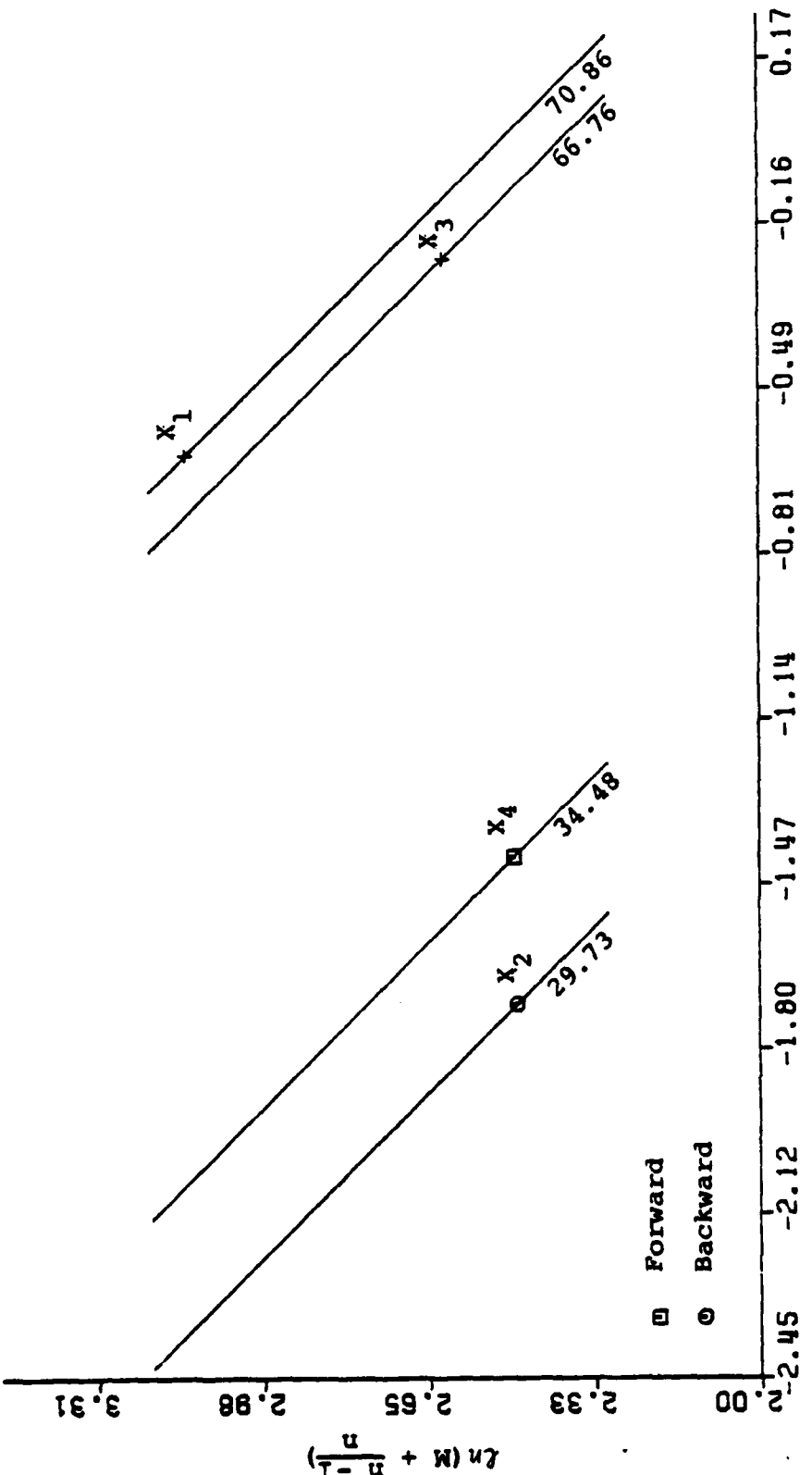


Figure 3. Graphic Display of Analogy Versus Fit. One-Variable Models.

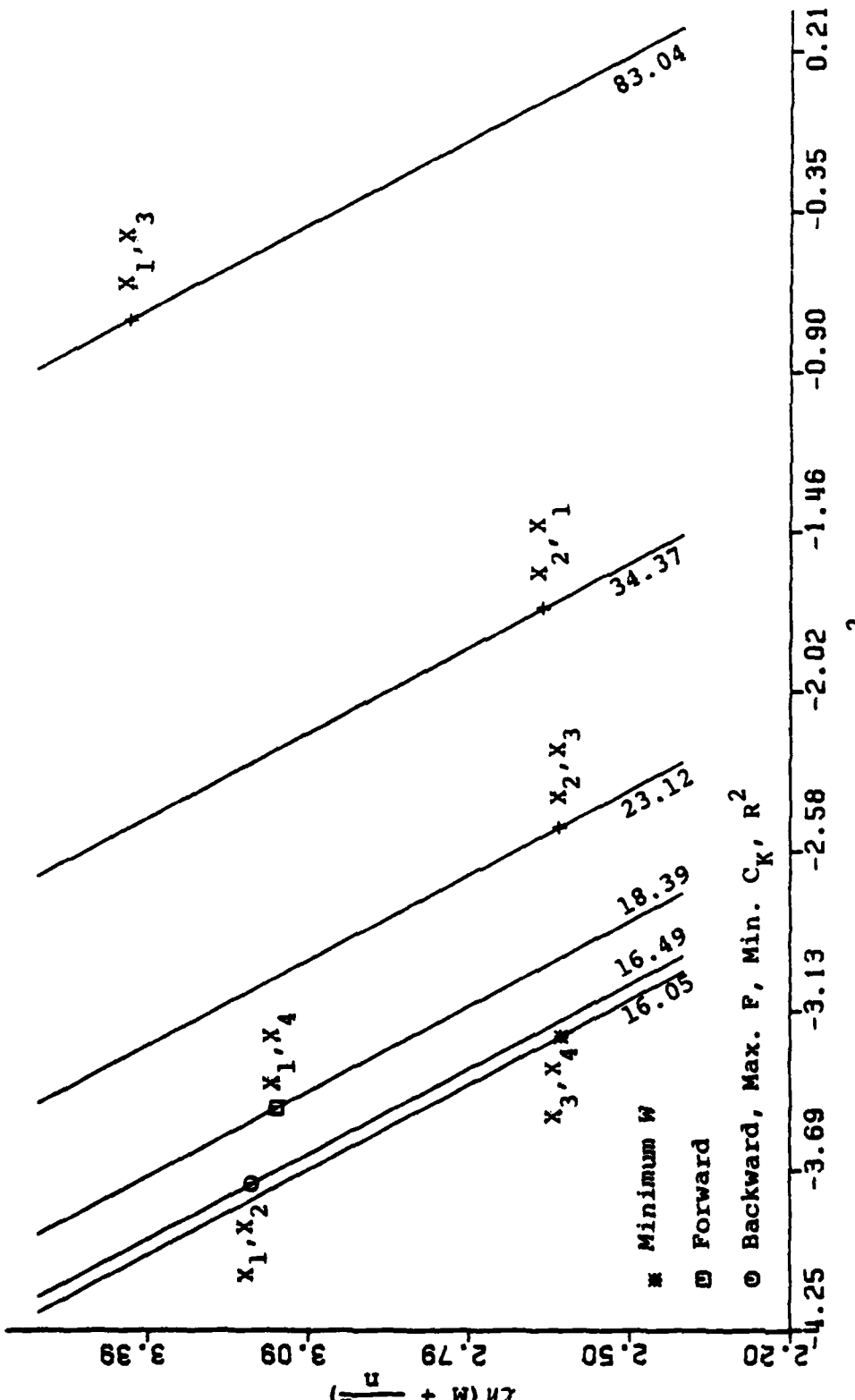


Figure 4. Graphic Display of Analogy Versus Fit. Two-Variable Models.

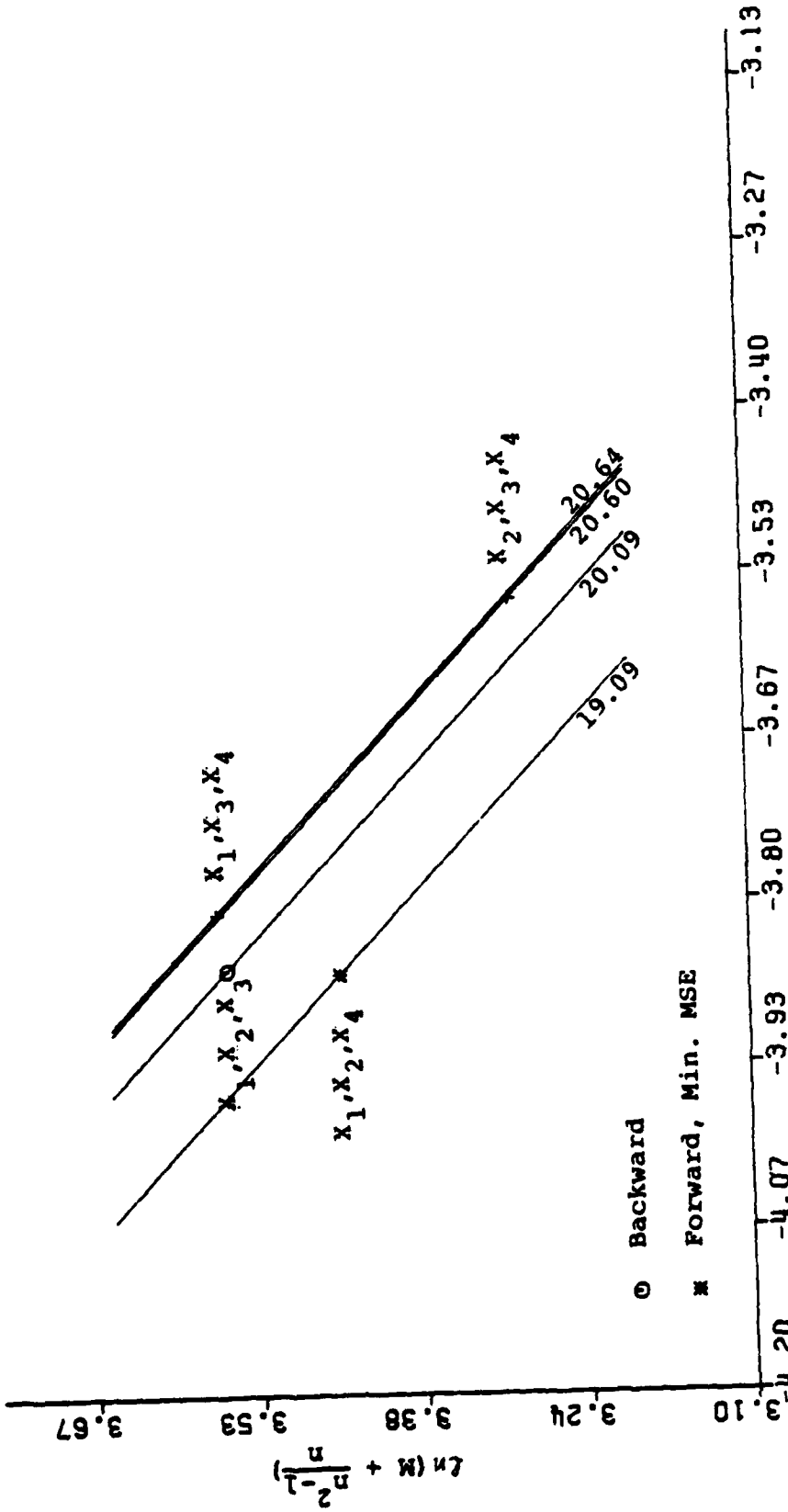


Figure 5. Graphic Display of Analogy Versus Fit. Three-Variable Models.

by any criterion are marked by a "+". For those which were selected, the legends indicate the corresponding criteria. (The models marked as FORWARD and BACKWARD will be discussed in the next chapter).

As can be seen from these displays, some of the two-variable models seem to be pointed out by several criteria. The model with variables X_1 and X_2 results in better fit of the data as its location in the graph indicates. However, the point under prediction is rather non-analogous to the data along these variables. Thus, the model with variables X_3 and X_4, although it is associated with a larger MSE (smaller R^2), results in a slightly smaller W. These two models and, perhaps, the one with variables X_1 and X_4 and the one with variables X_1, X_2 and X_4 would be the ones passing the first screening and scrutinized further.

For a second example, the data on page 352 in Draper and Smith (10) were used. They consist of twenty-five observations on nine predictor variables. The response measures the pounds of steam used monthly in a glycerine producing operation. The eighth row was set aside and predicted. This row was selected since it has been discussed in (1). A preliminary investigation suggested that variables X_3 and X_5 were not important carriers of information in the sense that they were not involved with any of the good models and they were the first ones to be eliminated by a BACKWARD procedure based on minimum reduction

in R^2. To reduce computation requirements and clutter in plots, they were not considered for further study. Figures 6-11 depict the situations. Because of the large number of models, only two "equi-W" lines were drawn for each graph for reference purposes across model sizes. They correspond to the smallest and the median W's and they are labeled by the width of the prediction interval.

This is a well behaved data set in that the models suggested by most criteria have the largest R^2 of their respective sizes. Also, as the graphs indicate, the point under prediction is rather analogous to the data along all dimensions (variables). The scales on the axes are the same on all graphs, making clear the general increase in the Mahalanobis distance as the number of variables increases. The model with variables X_2, X_4, X_6, X_8, X_9 and X_{10}, as these are labeled in Draper and Smith, was selected by the minimum W, the minimum MSE and the minimum C_k criteria. Notice that the model suggested by the Mean Square Error of Prediction criterion (variables X_4 and X_7) seems to be unacceptable on all other counts. It provides a $R^2 = 0.423$, which is very small compared with those of many other models. As a result, the prediction interval associated with it is very wide. This should underscore the fact that variable selection criteria are not universally applicable and can often lead to models which a careful analysis may find unacceptable. Therefore, they should be used with prudence and be accompanied by a careful analysis of the models they

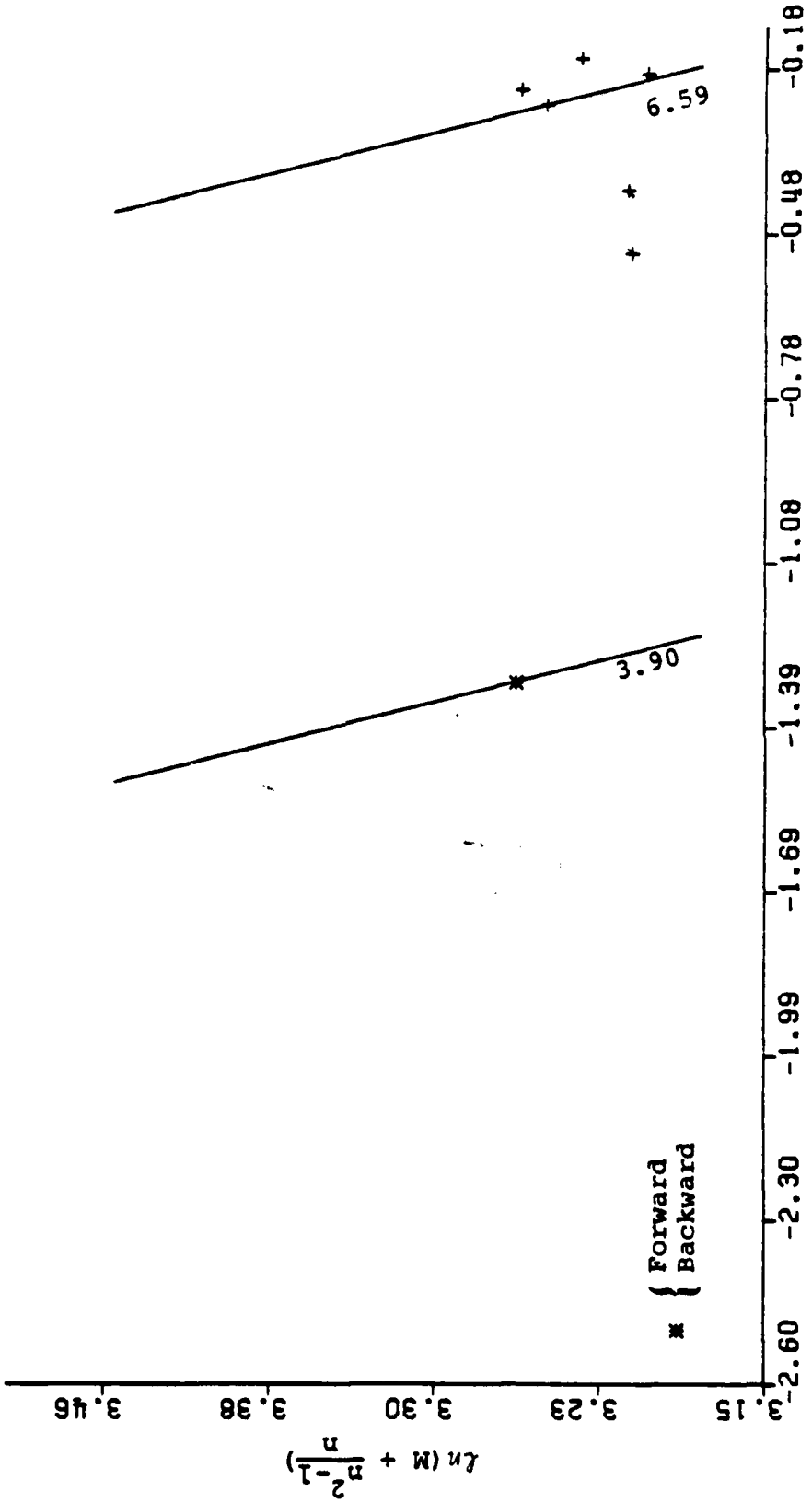


Figure 6. Graphic Display of Analogy Versus Fit. One-Variable Models.

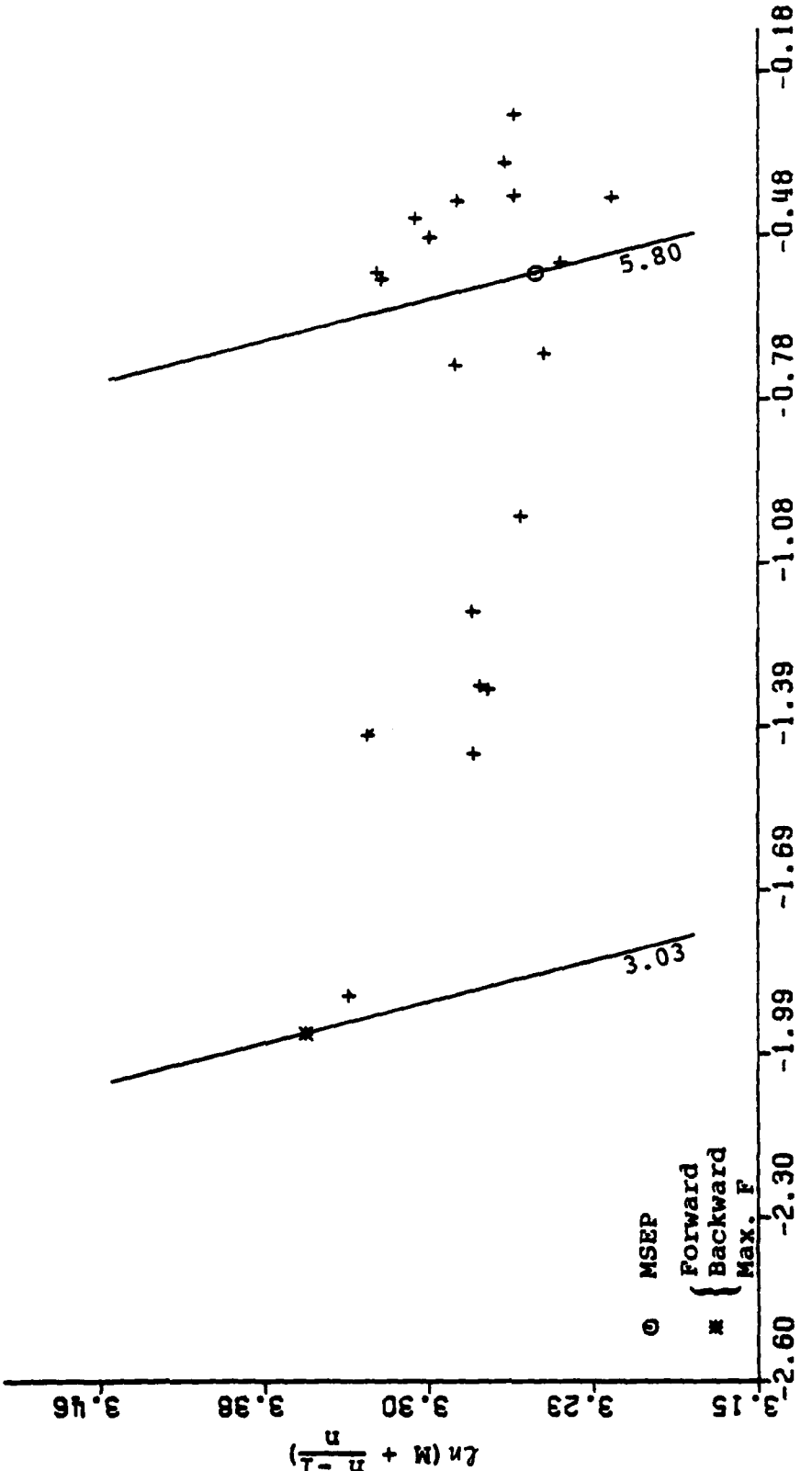


Figure 7. Graphic Display of Analogy Versus Fit. Two-Variable Models.

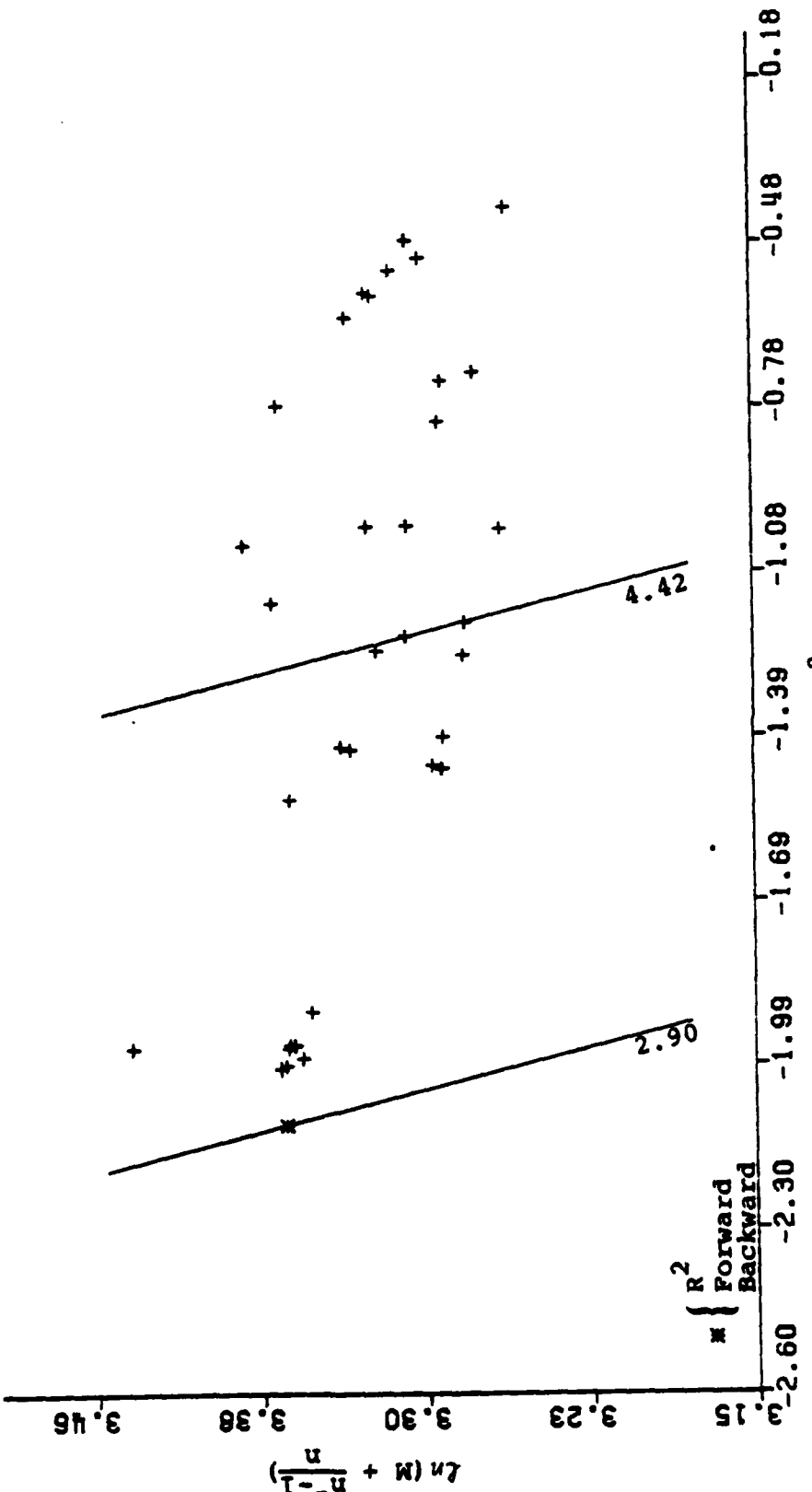


Figure 8. Graphic Display of Analogy Versus Fit. Three-Variable Models.

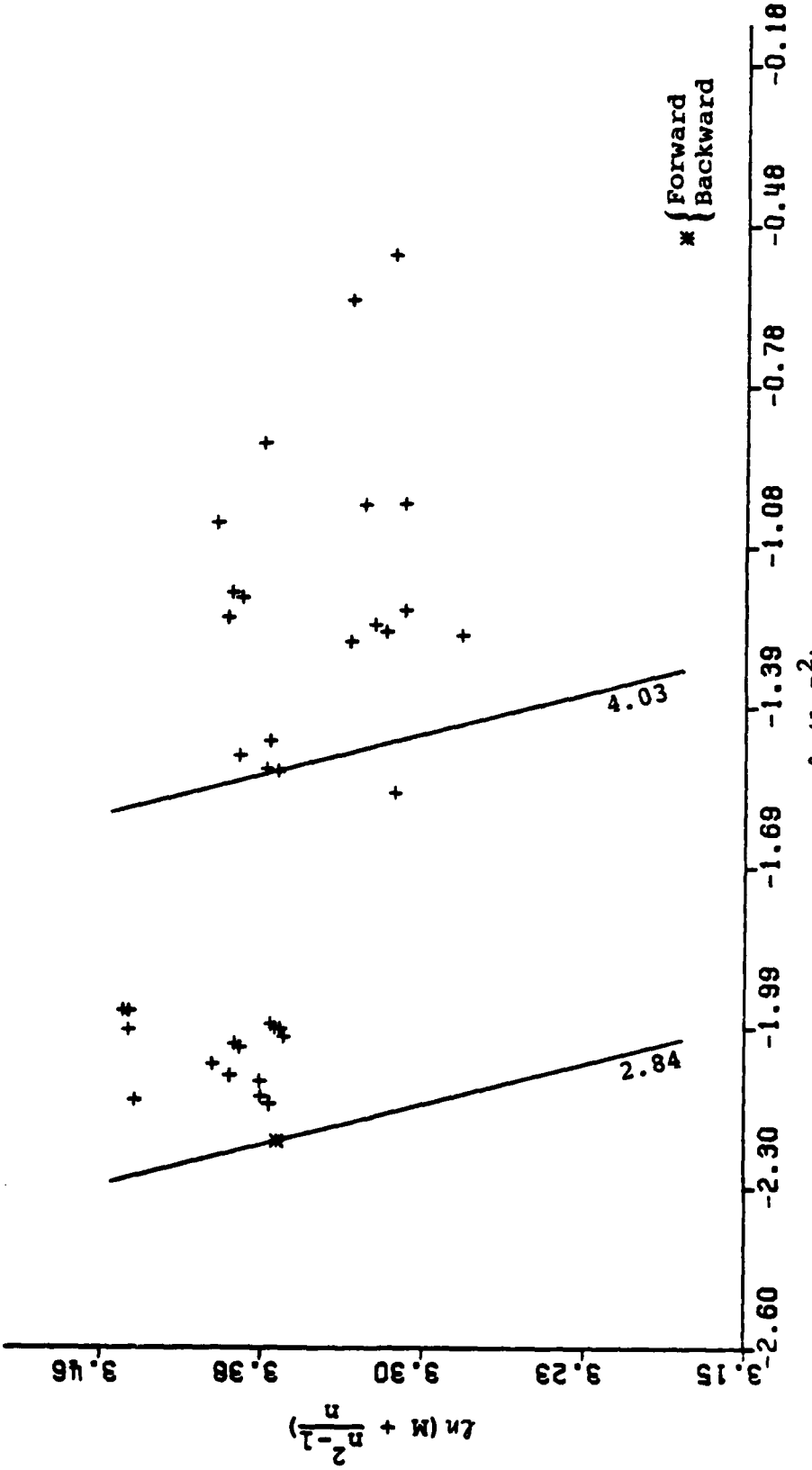


Figure 9. Graphic Display of Analogy Versus Fit. Four-Variable Models.

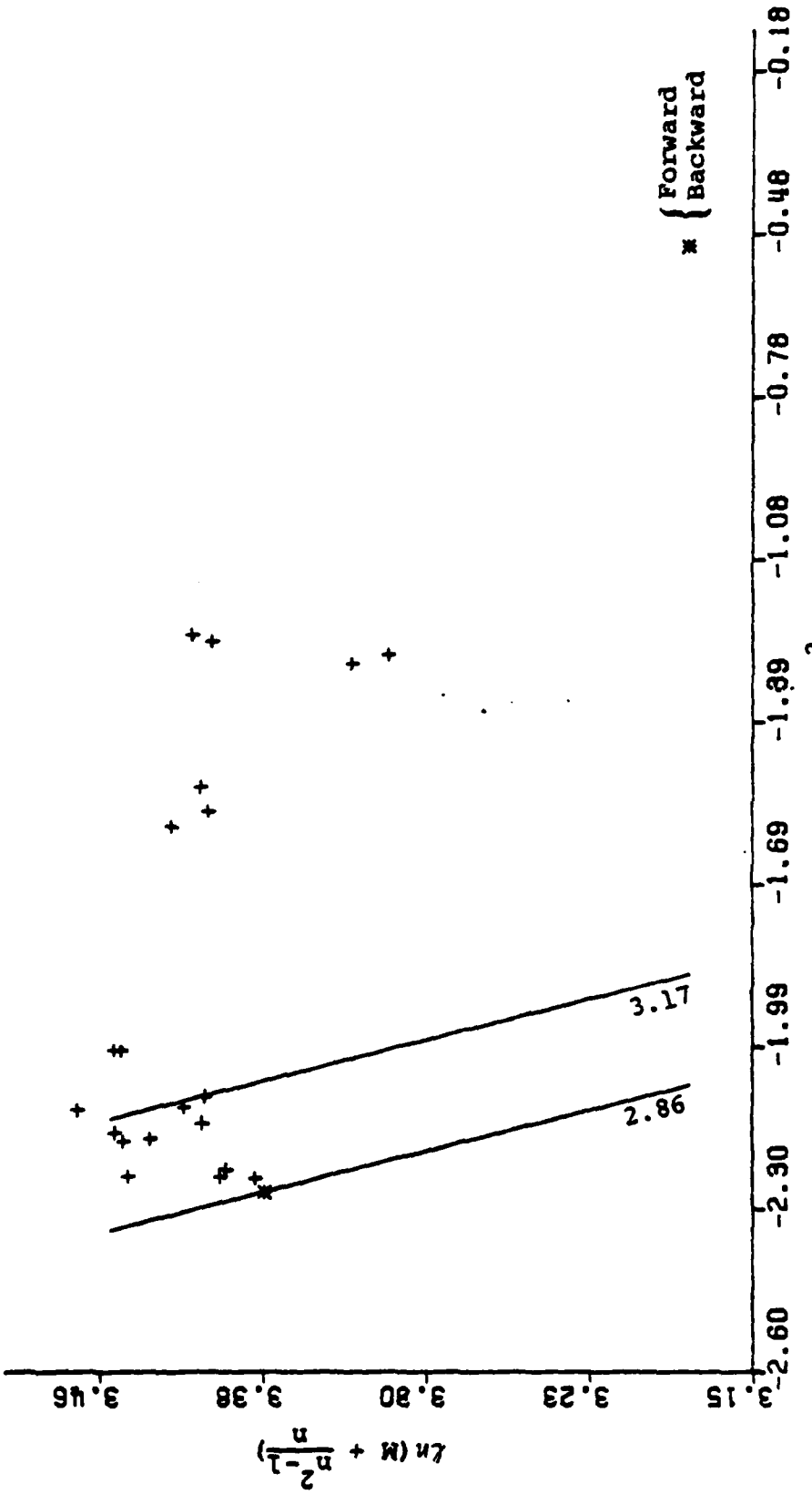


Figure 10. Graphic Display of Analogy Versus Fit. Five-Variable Models.

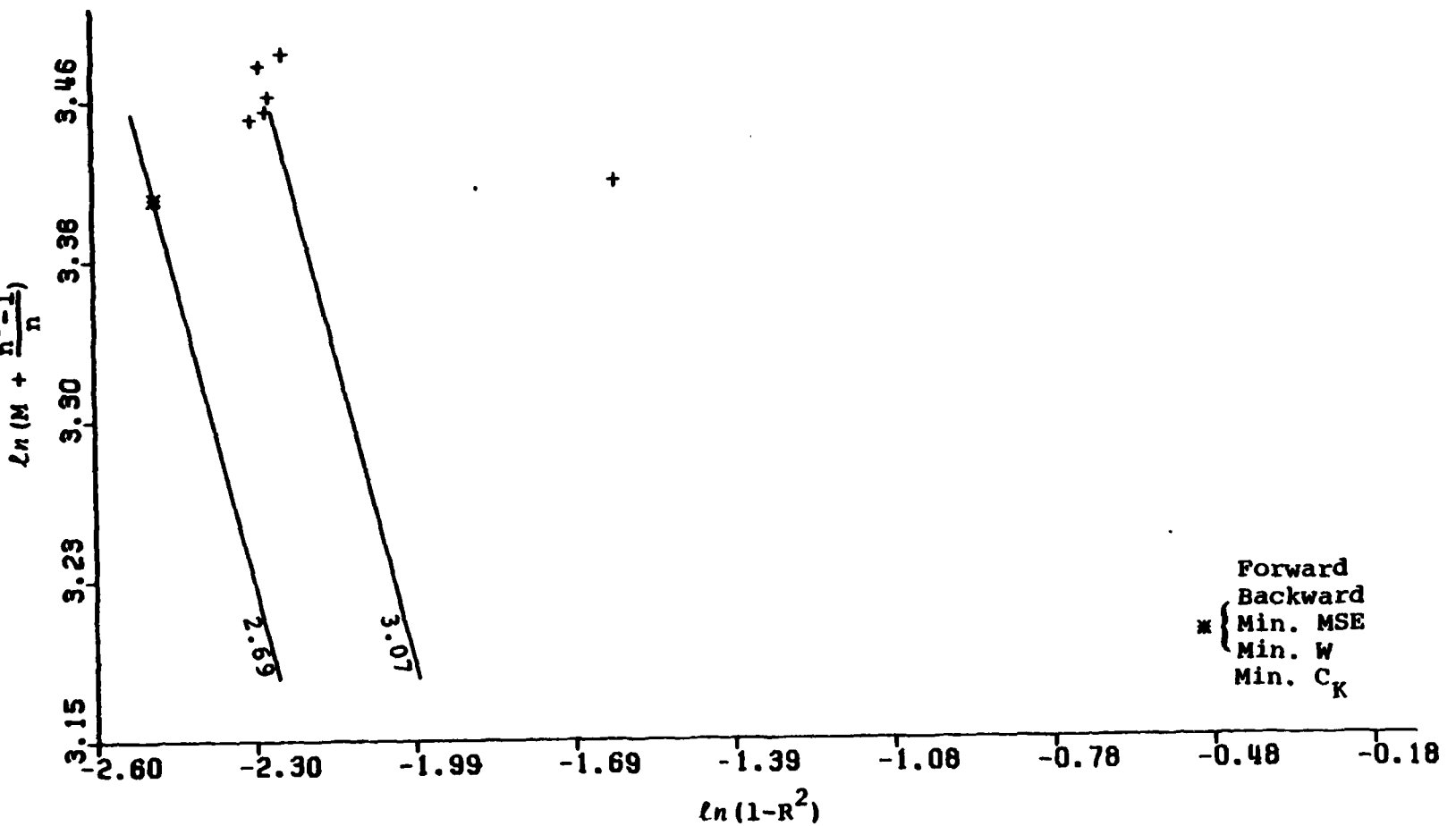


Figure 11. Graphic Display of Analogy Versus Fit. Six-Variable Models.

suggest along more than one lines. Model building should not be reduced to a mechanical selection of variables by any criterion.

In this chapter, the W criterion was developed and discussed. Its intuitive appeal in the specific problem of prediction at a known point is based on the fact that it has the potential of focusing attention to an aspect of the problem which could have been ignored otherwise. As is the case with every data analytic technique, the art of using this criterion to advantage must be developed through experience. In subsequent chapters, this methodology will be applied to real data sets in an effort to understand its properties. This should help in learning how to exploit its strong points and avoid its weaknesses.

CHAPTER IV

COMPUTATION

A Branch and Bound Algorithm

The need for the selection of a subset of variables becomes more imperative as the number p of potential predictors becomes large. At the same time, since the computing time needed for a search of all 2^p-1 possible regressions increases exponentially in p, it is clear that, for large p, a full search may be well beyond the budget considerations of the analysis. An algorithm which will identify the good models without actually performing all 2^p-1 regressions is, in such cases, highly desirable. Such algorithms exist for criteria which are simple functions of the sum of squared errors. The most efficient of these take advantage of the fact that the sum of squared errors associated with a model is a lower bound on the sums of squared errors of its submodels. In 1974, Furnival and Wilson (12) suggested a branch-and-bound algorithm whose efficiency is enhanced by the fact that the search is made by a simultaneous traversing of two trees, one for bounds and one for regressions. A semi-SWEEP operator is employed for the entering or removing of variables and the matrices needed at each stage are available from previous SWEEP's. This algorithm is the most efficient one known to date and problems involving 30 variables are well within its reach.

An attractive feature of this algorithm is that the "best" m models for each subset size k can be output without great loss of efficiency.

The technique proposed in the previous chapter is not simply related to the sum of squared errors, since it involves the coordinates of $\underline{x}$. Thus, for a given model size k, the model with smallest sum of squared errors may not yield the smallest W. This implies that the bounds utilized in Furnival's algorithm cannot be used for a similar search for models with small W. Somewhat less sharp bounds can nevertheless be obtained so that, with minor modifications, Furnival's approach can be adapted to the case at hand.

A univariate one-sample t^2 statistic is defined by $t^2 = n(x-\overline{X})^2/s^2$, where $\overline{X}$ is the mean of a sample of size n on a variable X, x is an independent observation on the same variable and s^2 is the sample variance of X. This is the exact univariate analog of Hotelling's one-sample T^2. Actually, $T^2 = n(\underline{x}-\underline{\overline{X}})S^{-1}(\underline{x}-\underline{\overline{X}})'$ is the square of the univariate t-ratio of that linear combination of the variables which inflates the t-ratio the most. (For a clear explanation of the derivation of T^2 see (17)). A univariate t-ratio corresponds to one such linear combination and therefore T^2 must be greater than or equal to the largest squared univariate t-ratio. All p univariate t-ratios can be computed and saved. Thus, a lower bound on T^2 associated with any set of X's is obtained. Since the Mahalanobis distance

$M = T^2/n$, a lower bound is also obtained on M. Recalling (3.11),

$$W = F_{1-\alpha;1,n-k-1} \frac{SSE}{n-k-1}\left[\frac{n+1}{n} + \frac{M}{n-1}\right],$$

if W_i^* denotes the smallest W currently available for models of size i at any stage of the search, then the submodels derived from a model of size k need not be examined if the sum of squared errors of that model is greater than:

$$\frac{n(n-1)(n-i-1)W_i^*}{[F_{1-\alpha;1,n-k-1}(n^2-1+t_{(i)}^2)]},$$

for all $i = 1,\ldots,k-1$, where $t_{(i)}^2$ is the i-th largest univariate t^2. Notice that this quantity needs to be calculated only when a model which improves W_i^* is encountered. For sharper bounds, this quantity can be recomputed at every stage using for $t_{(i)}^2$ the i-th largest univariate t^2 among the variables in the model under consideration.

Empirical experience with this algorithm suggests that it is approximately 5-10% less efficient when applied to W than when it is applied to other criteria which are simple functions of the sum of squared errors. As noted by Furnival, time requirements are heavily data dependent. This observation obviously applies to the W criterion as well, in which case efficiency will also depend on the value of $\underline{x}$. The importance of this last dependency diminishes for large n. In general, the efficiency of this algorithm when applied to the W criterion should be such that problems

of comparable size can be handled without much additional investment in computing time.

A Stepwise Algorithm

It is often the case that a suboptimal stepwise search is used in lieu of computing all regressions, especially in an early screening of a very large number of variables. Various stepwise procedures have been widely used for this purpose, all of which are variations of FORWARD selection and BACKWARD elimination (see e.g. (10)). These techniques, based on the SWEEP operator (29), were designed for criteria which are simple functions of the residual sum of squares and, hence, must be modified to deal with the W criterion. The computational complications associated with the W criterion can be overcome by exploiting the monotonicity property of Theorem 3.1. The SWEEP operator must be briefly considered first, in order to locate the quantities needed for a FORWARD selection and a BACKWARD elimination algorithm based on the minimum W criterion.

Given an originally symmetric positive definite matrix A, the SWEEP operator applied to the k-th diagonal element of A is defined as follows:

Step 1: Let $D = a_{kk}$.

Step 2: Divide row k by D.

Step 3: For every other row $i \neq k$, let $B = a_{ik}$.
Subtract B × row k from row i. Set $a_{ik} = -B/D$.

Step 4: Set $a_{kk} = 1/D$.

If a SWEEP is performed on diagonal element i, variable X_i is added to the current regression equation unless X_i is already in the regression in which case X_i is removed. Observe that the result of a SWEEP is an absolutely symmetric matrix, i.e., a matrix such that $a_{ij} = a_{ji}$ if an equal number of SWEEP's have been performed on elements i and j, and $a_{ij} = -a_{ji}$ otherwise. Thus, only the upper triangular part of A need be computed.

For the purposes of the W criterion, A must be set initially to the corrected sums-of-squares-cross-products matrix of the data, i.e.,

$$A = \begin{bmatrix} \underline{\underline{X}}'\underline{\underline{X}} & \underline{\underline{X}}'\underline{Y} \\ \underline{Y}'\underline{\underline{X}} & \underline{Y}'\underline{Y} \end{bmatrix}$$

where $\underline{\underline{X}}$ in this section will denote the original matrix $\underline{\underline{X}}$ corrected for the means $\bar{X}_1,\ldots,\bar{X}_p$ and $\underline{Y}$ will denote the original vector $\underline{Y}$ corrected for the mean $\bar{Y}$. In more familiar statistical terms, the matrix A is the covariance matrix of the data multiplied by (n-1). Variables are entered and deleted by sweeping on the corresponding diagonal elements of A. After each SWEEP, statistics on the regression of Y on the variables which have been swept in and submodel information are available. To illustrate, suppose

$$A = \begin{bmatrix} \underline{\underline{X}}_1'\underline{\underline{X}}_1 & \underline{\underline{X}}_1'\underline{\underline{X}}_2 & \underline{\underline{X}}_1'\underline{Y} \\ \underline{\underline{X}}_2'\underline{\underline{X}}_1 & \underline{\underline{X}}_2'\underline{\underline{X}}_2 & \underline{\underline{X}}_2'\underline{Y} \\ \underline{Y}'\underline{\underline{X}}_1 & \underline{Y}'\underline{\underline{X}}_2 & \underline{Y}'\underline{Y} \end{bmatrix}$$

where $\underset{\sim}{X}_1$ contains some of the columns of the data matrix $\underset{\sim}{X}$. Sweeping A on the diagonal elements of $\underset{\sim}{X}_1'\underset{\sim}{X}_1$ yields:

$$B = \begin{bmatrix} (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1} & (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'\underset{\sim}{X}_2 & (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'\underline{Y} \\ -\underset{\sim}{X}_2'\underset{\sim}{X}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1} & \underset{\sim}{X}_2'M_1\underset{\sim}{X}_2 & \underset{\sim}{X}_2'M_1\underline{Y} \\ -\underline{Y}'\underset{\sim}{X}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1} & \underline{Y}'M_1\underset{\sim}{X}_2 & \underline{Y}'M_1\underline{Y} \end{bmatrix}$$

where $M_1 = I-\underset{\sim}{X}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'$.

The rightmost column contains the regression coefficients and the sum of square errors of the regression of Y on the variables contained in $\underset{\sim}{X}_1$. The part $(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'\underset{\sim}{X}_2$ gives the coefficients of the regressions of each variable in $\underset{\sim}{X}_2$ on the variables in $\underset{\sim}{X}_1$, and the diagonal elements of $\underset{\sim}{X}_2'M_1\underset{\sim}{X}_2$ give the sum of square errors of the same regressions.

For a FORWARD selection algorithm, the reduction in SSE and the increase in the Mahalanobis distance resulting from the inclusion of a variable among those in $\underset{\sim}{X}_2$ can easily be computed. To illustrate, suppose that variables $X_1,\ldots,X_k$ have already been swept into the regression. The quantities needed for the computation of the new W, say W_j, resulting from the inclusion of variable X_j, $j = k+1,\ldots,p$, are:

$$\hat{x}_j = \bar{Y} - \sum_{i=1}^{k} B_{ij}\bar{X}_i + \sum_{i=1}^{k} B_{ij}x_i \,, \tag{4.1}$$

$$S_{jj}(1-r^2) = B_{jj}/(n-1) \tag{4.2}$$

and

$$SSE_j = B_{(p+1)(p+1)} - B^2_{j(k+1)}/B_{jj} \qquad (4.3)$$

where $S_{jj}(1-r^2)$ denotes, as in Chapter III, the conditional variance of X_j on $X_1,\ldots,X_k$. Therefore, using the monotonicity property of Theorem 3.1, the new W resulting from the introduction of variable X_j into the regression equation is

$$W_j = F_{1-\alpha;1,n-k-1}\frac{SSE_j}{n-k-2}\left[\frac{n+1}{n} + \frac{M}{n-1} + \frac{(X_j-\hat{X}_j)^2}{B_{jj}}\right] \qquad (4.4)$$

where M is the Mahalanobis distance associated with variables $X_1,\ldots,X_k$. Thus, variable X_j is the next variable to sweep into the regression, if $W_j = \min\{W_i,\ i = k+1,\ldots,p\}$.

For a BACKWARD elimination algorithm, the variable to be swept out is determined as follows:
If variable X_j, $j = 1,\ldots,k$ is deleted, the error sum of squares becomes

$$SSE_j = B_{(p+1)(p+1)} + B^2_{j(p+1)}/B_{jj}. \qquad (4.5)$$

For the new Mahalanobis distance, the coefficients of the regression of X_j on X_ℓ, $\ell = 1,\ldots,k$, $\ell \neq j$ need to be computed. These are given by

$$C_i = -B_{ij}/B_{jj}, \quad i = 1,\ldots,k,\ i \neq j. \qquad (4.6)$$

Thus,

$$\hat{x}_j = \bar{Y} - \sum_{\substack{i=1\\ i\neq j}}^{k} C_i\bar{X}_i + \sum_{\substack{i=1\\ i\neq j}}^{k} C_iX_i.$$

Again using Theorem 3.1, W becomes

$$W_j = F_{1-\alpha;1,n-k} \frac{SSE_j}{n-k} \left[\frac{n+1}{n} + \frac{M}{n-1} - \frac{(X_j-\hat{X}_j)^2 B_{jj}}{(n-1)^2}\right] \quad (4.7)$$

where M is as noted above. Thus, variable X_j is the next variable to be swept out, if $W_j = \min\{W_i,\ i = 1,\ldots,k\}$.

Given the monotonicity result of Chapter III, in FORWARD selection the computed M is added to the current Mahalanobis distance, while in BACKWARD elimination it is subtracted. The current Mahalanobis distance must be saved at each stage.

The computing time requirements for such algorithms pose no limitations on their applicability in problems of sizes normally encountered in practice. At issue is the degree to which the models selected for each subset size differ from the ones found optimal by the criterion employed when a full search is done. In order to gain some insight into this, the two stepwise procedures were applied to the data sets used in Chapter III. The models selected by the FORWARD selection and the ones selected by the BACKWARD elimination are shown in Figures 3-11. In the four variable example, the algorithms identified the better models for each subset size. The BACKWARD procedure identified the best one-variable model, while the FORWARD procedure located the best three variables. They both missed the overall optimal model (X_3,X_4), however, the two-variable model selected by BACKWARD elimination had a W very close to the optimal one.

In the second example, depicted in Figures 6-11, the two procedures selected the same model for all model sizes. Observe that, in all cases, the models selected by these sub-optimal procedures coincided with the minimum-W-optimal ones. This, of course, is the most desirable situation. The degree to which it will happen in practice depends on the particular set of data. However, if these two examples offer any indication, it seems that the stepwise procedures can fruitfully be employed either in thinning down a large number of variables to a subset on which a full search by means of the branch-and-bound algorithm will be economically feasible, or by themselves. The better models of each subset size should be identified at least in the cases of well behaved data sets.

CHAPTER V

ON THE INFLUENCE OF OBSERVATIONS ON W

Theory and Discussion

The influence of individual observations on the various quantities of interest in a statistical analysis of data has received considerable attention in the recent literature (8), (13), (18), etc. It is argued that observations which significantly affect (have high leverage on) such quantities ought to be given careful scrutiny. The object is to detect "outlying" points and to investigate them further, examining whether the analysis can be enhanced by setting them aside. Possible errors of transcription, for instance, might be discovered. More realistically, in cases of designed experiments, such knowledge may prove useful in suggesting ways in which the design may be improved. Taking more measurements in the space of the explanatory variables could improve the analysis.

Identification of outliers does not necessarily imply, or argue for, the rejection of such points. It is only meant as a tool for the analysis of data and should be used with caution. Nevertheless, the inclusion of faulty data can adversely affect the analysis to a substantial degree. This point has recently received attention in the literature. Hoaglin and Welsch (18) studied the "hat" matrix

$\underline{\underline{X}}(\underline{\underline{X}}'\underline{\underline{X}})^{-1}\underline{\underline{X}}'$. They suggested an approach combining the information carried by the hat matrix and the studentized residuals in an effort to discover exceptional and/or discrepant points.

Cook (7) proposed the distance

$$D_i = [(\underline{b}-\underline{b}_{(i)})'\underline{\underline{X}}'\underline{\underline{X}}(\underline{b}-\underline{b}_{(i)})]/(k\times MSE) \ ,$$

where $\underline{b}_{(i)}$ denotes the estimated coefficients obtained without observation i, as a measure of the influence of the i-th data point. He, too, related such influences to the hat matrix, the studentized residuals and residual variances.

Welsch and Peters (36) suggested methods for examining more than two observations at a time and placed emphasis on the computational aspects of these diagnostic measures.

Gentleman and Wilk (13) developed analysis of variance methods to identify outlying subsets of K observations.

The investigations above are mainly concerned with the influence of outliers on the parameter estimates rather than prediction. In the context of this investigation, an observation (or group of observations) may be termed exceptional if W changes significantly when that observation (or group of observations) is set aside and the least squares calculations are performed on the reduced data set. The effected change in W will be investigated by means of the ratio of prediction variances. Some new notation will facilitate the exposition of this chapter.

Partition the matrix $\underset{\sim}{X}$ as $\{\underset{\sim}{X}_1', \underset{\sim}{X}_2'\}'$, where $\underset{\sim}{X}_1$ is $n_1 \times p$ and $\underset{\sim}{X}_2$ is $n_2 \times p$ with $n_1 + n_2 = n$. The vector $\underset{\sim}{Y}$ is partitioned in like manner into components $\underset{\sim}{Y}_1$ and $\underset{\sim}{Y}_2$. Without loss of generality, assume that the observations in $(\underset{\sim}{X}_2, \underset{\sim}{Y}_2)$ are set aside. Let s^2 and s_1^2 denote the mean square errors of the full model and the submodel, respectively. A superscript (1) will indicate that the quantity to which it is attached has been computed from the regression using $\underset{\sim}{X}_1$ and $\underset{\sim}{Y}_1$ only. Let $\underset{\sim}{e}_1$ and $\underset{\sim}{e}_2$ be the $n_1 \times 1$ and $n_2 \times 1$ vectors of residuals corresponding to $\underset{\sim}{Y}_1$ and $\underset{\sim}{Y}_2$ respectively when the full data base is used. In accord with the convention stated above, then $\underset{\sim}{e}_1^{(1)}$ and $\underset{\sim}{e}_2^{(1)}$ would be the vectors of residuals corresponding to $\underset{\sim}{Y}_1$ and $\underset{\sim}{Y}_2$ resulting from fitting the model to $\underset{\sim}{X}_1$ and $\underset{\sim}{Y}_1$. A well known result (for example see Bingham (4)) yields

$$s_1^2 = \frac{(n-p)s^2 - \underset{\sim}{e}_2'[I - \underset{\sim}{X}_2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}_2']^{-1}\underset{\sim}{e}_2}{(n_1-p)} \tag{5.1}$$

where I denotes the $n_2 \times n_2$ identity matrix. Letting

$$\gamma^2 = \frac{s^2[1 + \underset{\sim}{x}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{x}']}{s_1^2[1 + \underset{\sim}{x}(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{x}']} \tag{5.2}$$

i.e., the ratio of prediction variances at $\underset{\sim}{x}$ and substituting (5.1) into (5.2), it is easy to show that

$$\gamma^2 = \frac{n_1 - p}{n-p} \times Q \times H, \tag{5.3}$$

where

$$Q = 1 + \frac{\underline{e}_2'[I-\underline{X}_2(\underline{X}'\underline{X})^{-1}\underline{X}_2']^{-1}\underline{e}_2}{(n_1-p)s_1^2}$$

and

$$H = \frac{[1+\underline{x}(\underline{X}'\underline{X})^{-1}\underline{x}']}{[1+\underline{x}(\underline{X}_1'\underline{X}_1)^{-1}\underline{x}']} .$$

Using another identity from (4), namely

$$\underline{e}_2 = [I+\underline{X}_2(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_2']^{-1}\underline{e}_2^{(1)} = [I-\underline{X}_2(\underline{X}'\underline{X})^{-1}\underline{X}_2']\underline{e}_2^{(1)} \tag{5.4}$$

γ^2 can also be written as

$$\gamma^2 = \frac{n_1-p}{n-p} \times Q^{(1)} \times H \tag{5.5}$$

where

$$Q^{(1)} = 1 + \frac{\underline{e}_2^{(1)'}[I+\underline{X}_2(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_2']^{-1}\underline{e}_2^{(1)}}{(n_1-p)s_1^2} .$$

Relations (5.3) and (5.5) express the ratio of prediction variances, γ^2, in terms of quantities which yield to intuitive interpretations. These quantities are studied next, in an effort to isolate the characteristics of observations whose deletion results in a significant change in W.

It is clear that a reduction in W is obtained if, and only if,

$$\gamma^2 > F_{1-\alpha;1,n_1-p}/F_{1-\alpha;1,n-p}.$$

Consider the three factors comprising γ^2. Clearly, $\{(n_1-p)/(n-p)\} < 1$. The following theorem shows that the last factor also has this property.

<u>Theorem 5.1</u>: $$\frac{1+\underline{x}(\underline{X}'\underline{X})^{-1}\underline{x}'}{1+\underline{x}(\underline{X}_1'\underline{X}_1)^{-1}\underline{x}'} \quad 1.$$

<u>Proof</u>: It suffices to prove the result for the case $n_2 = 1$. Since $\underline{X}'\underline{X} = \underline{X}_1'\underline{X}_1 + \underline{X}_2'\underline{X}_2$ and

$$(\underline{X}'\underline{X})^{-1} = (\underline{X}_1'\underline{X}_1)^{-1} - \frac{(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_2'\underline{X}_2(\underline{X}_1'\underline{X}_1)^{-1}}{1+\underline{X}_2(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_2'}$$

it follows that,

$$\underline{x}(\underline{X}_1'\underline{X}_1)^{-1}\underline{x}' = \underline{x}(\underline{X}'\underline{X})^{-1}\underline{x}' + \frac{[\underline{x}(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_2']^2}{1+\underline{X}_2(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_2'}. \qquad (5.6)$$

Since $(\underline{X}_1'\underline{X}_1)^{-1}$ is positive definite, the second term in the right hand side of (5.6) is also positive, completing the proof.

The second factor in (5.5) is greater than one. This is so because the matrix $I+\underline{X}_2(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_2'$ is positive definite, being the sum of positive definite matrices and, therefore, $[I+\underline{X}_2(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_2']^{-1}$ is also positive definite. Since only the last two factors in γ^2 depend on the components of $\underline{X}_2$, they will expose the characteristics of observations which affect W. The factor

$$t_2^2 = \frac{\underline{e}_2^{(1)\prime}[I+\underline{X}_2(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_2']^{-1}\underline{e}_2^{(1)}}{s_1^2}$$

is studied first.

Observe that $|I+\underline{\underline{X}}_2(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{\underline{X}}_2'|$ is a measure of the collective distance of the points in $\underline{\underline{X}}_2$ from the rest of the data. This is meant in the following sense: Clearly, if $n_2 = 1$, then

$$I+\underline{\underline{X}}_2(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{\underline{X}}_2' = \frac{n+1}{n} + \frac{M}{n-1}$$

where M is the Mahalanobis distance of the point $\underline{\underline{X}}_2$ from the data base $\underline{\underline{X}}_1$. When $n_2 > 1$, a large $|I+\underline{\underline{X}}_2(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{\underline{X}}_2'|$ will indicate that the points in $\underline{\underline{X}}_2$ are either far from the centroid of $\underline{\underline{X}}_1$, or that their covariance structure is different from that of $\underline{\underline{X}}_1$, or both. Thus, for a reduction in W, the determinant of $I+\underline{\underline{X}}_2(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{\underline{X}}_2'$ must be small. That is to say, other things being equal in (5.5), points near the centroid of $\underline{\underline{X}}_1$ are more likely to cause W to decrease when they are omitted. This should be intuitively appealing. For the variances of the estimated coefficients to be small, the data points must be widely dispersed *in* X-space.

Next, observe that the residuals $\underline{e}_2^{(1)}$ must be large in absolute value. In other words, the Y values of the observations to be set aside must be discrepant in the sense that, when the model is built on the remaining n_1 observations, $(\underline{\underline{X}}_2, \underline{Y}_2)$ are not fit well. Cook (8), Hoaglin and Welsch (18) and others have linked the residual t_2^2 with the influence of the set $(\underline{\underline{X}}_2, \underline{Y}_2)$ on the coefficient estimates. Also, as should be obvious, s_1^2 should be small. For diagnostic purposes, it will be more convenient to look at the

factors comprising t_2^2 simultaneously. Notice that $s_1^2[I+\underset{\sim}{X}_2(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_2']$ is the usual estimate of the covariance matrix of the residuals $\underset{\sim}{e}_2^{(1)}$ (which, incidentally, are the basis for Allen's PRESS criterion). Therefore, t_2^2 can be viewed as a collective studentized residual corresponding to the omitted set $(\underset{\sim}{X}_2, \underset{\sim}{Y}_2)$. In fact, when $n_2 = 1$, t_2^2 reduces to

$$\frac{\underset{\sim}{e}_2^{(1)} \cdot \underset{\sim}{e}_2^{(1)}}{s_1^2[1+\underset{\sim}{X}_2(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_2'}$$

which is exactly what is called the studentized residual. When the rows in $\underset{\sim}{X}_2$ have been specified in advance, the quantity

$$t_2^2/n_2 = \frac{\underset{\sim}{e}_2^{(1)} \cdot [I+\underset{\sim}{X}_2(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_2']^{-1}\underset{\sim}{e}_2^{(1)}}{n_2 s_1^2}$$

is distributed as F with n_2 and n_1-p degrees of freedom since the numerator is distributed as $\sigma^2\chi^2(n_2)/n_2$, independently of the denominator which is distributed as $\sigma^2\chi^2(n_1-p)/(n_1-p)$. Thus, observations whose collective studentized residual is significantly large ought to be investigated further. It should be noted that t_2 depends not only on the individual residuals, but on their correlations as well. Although in practice observations whose studentized residuals are small rarely reduce W significantly when they are combined with others, this is not always the case. Cases have been observed where the pair which causes W to decrease the most consists of observations which, if deleted

individually, would cause W to increase. To such a "masking" effect the last factor in γ^2 may contribute significantly. One way to look at

$$\frac{1+\underline{x}(\underline{\underline{X}}'\underline{\underline{X}})^{-1}\underline{x}'}{1+\underline{x}(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{x}'}$$

is as a measure of the relative distances from $\underline{x}$ to $\underline{\underline{X}}$ and $\underline{\underline{X}}_1$ respectively. As noted above, this factor cannot be greater than one. For a maximum reduction in W, it should be as close to one as possible. This would imply that the deletion of $\underline{\underline{X}}_2$ does not greatly increase the Mahalanobis distance from $\underline{x}$ to the data base. This factor can also be studied in terms of residual correlations. Notice that

$$1-\eta^2 = 1 - \frac{[1+\underline{x}(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{x}']-[1+\underline{x}(\underline{\underline{X}}'\underline{\underline{X}})^{-1}\underline{x}']}{1+\underline{x}(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{x}'} \qquad (5.7)$$

Recall that the residuals $y-\hat{y}^{(1)}$ and $Y_2-\hat{Y}_2^{(1)}$ have a sampling distribution which, under the usual assumptions on ε, is normal with mean vector $\underline{0}$ and covariance matrix

$$C = \sigma^2 \begin{bmatrix} 1+\underline{x}(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{x}' & x(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{\underline{X}}_2' \\ \underline{\underline{X}}_2(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{x}' & 1+\underline{\underline{X}}_2(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{\underline{X}}_2' \end{bmatrix}.$$

Simple algebra will show that η^2 is the square of the multiple correlation between $y-\hat{y}^{(1)}$ and $Y_2-\hat{Y}_2^{(1)}$. This observation allows the following intuitively obvious statement: For a reduction in W, the deleted observations $(\underline{\underline{X}}_2, \underline{Y}_2)$ should not

contribute significantly in the explanation of the variability in y, beyond that which is provided by the retained rows $(\underline{X}_1, \underline{Y}_1)$.

It may be of interest to note the similarity between row deletion and variable selection. In the latter, the $\binom{p}{1}+\binom{p}{2}+\ldots+\binom{p}{p}$, $p < n$ submodels are investigated, in order to find the minimal, in some sense, subset of variables which adequately explains the data. In the former, the model is kept fixed and the possible $\binom{n}{1}+\binom{n}{2}+\ldots+\binom{n}{n_2}$, $n_2 << (n-p)$ data subsets are explored in order to find the maximal set which is adquately explained by the model. The postulated model form is held fixed, as the notion of "outlier" is valid only relative to a prespecified model form. The notation above indicates that n_2 must be small relative to n-p, in order to have a sufficient number of error degrees of freedom left. Notice that, as $n_2 \to n-p$, the sum of the squares of the residuals approaches zero, thus creating a false sense of security. In practice, if observations were to be deleted one at a time as long as some measure of fit, or W, "improved", most of the time all error degrees of freedom would be exhausted. This can be seen as follows: Observe first that the hat matrix $H = \underline{X}(\underline{X}'\underline{X})^{-1}\underline{X}'$ is a projection matrix, i.e., $H \times H = H$, and as such, it has all its eigenvalues equal to zero or one ((16), Thrm. 1.7.2, p. 39). The number of nonzero eigenvalues is equal to the rank of H. In the full rank case,

rank(H) = rank ($\underset{\sim}{X}$) = p. Hence, trace(H) = p, i.e., $\sum_{i=1}^{n} h_{ii} = p$. Also, since the ratio of the mean square error of the full model to that of the reduced one is equal to $[(n-p-1)/(n-p)][1+t_2^2/(n-p-1)]$, it follows that the mean square error will decrease if and only if an observation with $t_2^2 > 1$ is deleted. Finally, let

$$t^2 = \frac{e_2^2}{s^2[1-\underset{\sim}{X}_2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}_2']} .$$

Then $t_2^2 > 1 \iff t^2 > 1$. These observations and the lemma which follows will help in proving Theorem 5.2 below.

Lemma 5.1. Let $z_1, z_2, \ldots, z_n$ be any set of non-negative numbers. Let $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$, and let $a_1, a_2, \ldots, a_n$ be another set of non-negative numbers such that $\sum_{i=1}^{n} a_i = n$. Then there exists i such that $z_i/(a_i\bar{z}) \geq 1$.

Proof: Suppose that $z_i/(a_i\bar{z}) < 1$ for all i. Then, $z_i < a_i\bar{z}$ for all $i \iff \sum_{i=1}^{n} z_i < \sum_{i=1}^{n} a_i\bar{z} \iff \sum_{i=1}^{n} z_i < \bar{z}\sum_{i=1}^{n} a_i$ $\iff \sum_{i=1}^{n} z_i < n\bar{z} \iff \sum_{i=1}^{n} z_i < \sum_{i=1}^{n} z_i$, which is obviously false.

Note: Either there exists an i such that $z_i/(a_i\bar{z}) > 1$, or $z_i/(a_i\bar{z}) = 1$ for all i.

Theorem 5.2. With probability one, there exists at least one observation which, if deleted, will cause the mean square error to decrease.

Proof: Equivalently, using the observations made above, there exists at least one i for which

$$t^2 = \frac{e_i^2}{s^2(1-h_{ii})} > 1.$$

Write

$$s^2(1-h_{ii}) = (1-h_{ii}) \sum_{i=1}^{n} \frac{e_i^2}{n-p}$$

$$= \frac{n(1-h_{ii})}{n-p} \frac{1}{n} \sum_{i=1}^{n} e_i^2 .$$

Notice that

$$\sum_{i=1}^{n} \frac{n(1-h_{ii})}{n-p} = \frac{n}{n-p} \sum_{i=1}^{n} (1-h_{ii})$$

$$= \frac{n}{n-p} (n-p) = n.$$

Now,

$$t^2 = \frac{e_i^2}{(1-h_{ii}) \frac{n}{n-p} \sum_{i=1}^{n} \frac{e_i^2}{n}}$$

and the claim follows from Lemma 5.1 by letting

$$z_i = e_i^2, \quad \bar{z} = \frac{1}{n} \sum_{i=1}^{n} e_i^2 \quad \text{and} \quad a_i = \frac{n(1-h_{ii})}{n-p} .$$

With respect to W, this result implies the following: Given that the factor $1-\eta^2$ is, for most observations, close to one, especially when n is large relative to p, there will

tend to exist at least one observation whose deletion from that data base will make $\gamma^2 > F_{1-\alpha;1,n_1-p}/F_{1-\alpha;1,n-p}$, thus causing W to decrease. In practice, this seems frequently to be the case. Therefore, a decrease in W should not be the objective in determining observations to be set aside. These results suggest that a maximal n_2 should be chosen in advance, according to the analyst's a priori belief about the maximum possible (or likely) number of outliers, and such that $n_2 << (n-p)$. Then, subsets of n_2 or fewer observations whose deletion greatly reduces W should be examined in view of the discussion in the beginning of this chapter. It should be reemphasized that the object of such analysis is not the rejection or observations, but rather the gaining of insight about the data under investigation. Points whose presence in the data base has a significant effect on the quantities of interest should be scrutinized. If the validity of such observations is beyond question, the reasons for such behavior should be investigated. This should be done in an effort to gain a more penetrating insight into the data under investigation which insight might suggest otherwise overlooked remedial action such as the need for collection of more data in certain regions of the explanatory variables' space, when that is possible.

Computation

For computational purposes it will be preferable to express γ^2 in terms of the full model. For this purpose,

t_2^2 can be written as

$$t_2^2 = (n_1-p)t^2/(n-p-n_2t^2) \tag{5.8}$$

where

$$t^2 = \frac{\underline{e}_2'[I-\underline{X}_2(\underline{X}'\underline{X})^{-1}\underline{X}_2']^{-1}\underline{e}_2}{n_2s^2} . \tag{5.9}$$

Observe that t^2 and t_2^2 are one-to-one monotone functions of each other. Using their relation, t^2 can be transformed to an F statistic. Also, using the identity found in (27) p. 29,

$$1-\eta^2 = [1+\underline{x}(\underline{X}'\underline{X})^{-1}\underline{x}'] \,/\, \{[1+\underline{x}(\underline{X}'\underline{X})^{-1}\underline{x}'] + \underline{x}(\underline{X}'\underline{X})^{-1}\underline{X}_2'[I-\underline{X}_2(\underline{X}'\underline{X})^{-1}\underline{X}_2']^{-1}\underline{X}_2(\underline{X}'\underline{X})^{-1}\underline{x}'\}. \tag{5.10}$$

An expression for γ^2 in terms of the full model is now obtained if (5.8) and (5.10) are substituted into

$$\gamma^2 = \frac{n_1-p}{n-p}\left[1 + \frac{t_2^2}{n_1-p}\right](1-\eta^2). \tag{5.11}$$

This formula can be used as a building block for a computational procedure. However, one should be aware of the computational instability which is inherent to the problem of deleting a row (or block of rows) from a regression. See Chambers (6) for a discussion on the subject. The symmetric matrix

$$A = \begin{bmatrix} \underline{X}'\underline{X} & \underline{X}'\underline{Y} & \underline{X}' & \underline{x}' \\ & \underline{Y}'\underline{Y} & \underline{Y}' & 0 \\ & & I & 0 \\ & & & -1 \end{bmatrix}$$

can be set up.

After sweeping on the diagonal elements of $\underline{X}'\underline{X}$,

$$B = \begin{bmatrix} (\underline{X}'\underline{X})^{-1} & (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y} & (\underline{X}'\underline{X})^{-1}\underline{X}' & (\underline{X}'\underline{X})^{-1}\underline{x}' \\ & \underline{Y}'\underline{Y}-\underline{Y}'\underline{X}(\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y} & \underline{e}' & -\hat{y} \\ & & I-\underline{X}(\underline{X}'\underline{X})^{-1}\underline{X}' & -\underline{X}(\underline{X}'\underline{X})^{-1}\underline{x}' \\ & & & -1-\underline{x}(\underline{X}'\underline{X})^{-1}\underline{x}' \end{bmatrix}$$

where $\underline{e}'$ denotes the $1\times n$ vector of residuals. If $n_2 = 1$, the quantities needed for γ^2 are available in B. If $n_2 > 1$, the matrix

$$C = \begin{bmatrix} I-\underline{X}_2(\underline{X}'\underline{X})^{-1}\underline{X}_2' & -\underline{X}_2(\underline{X}'\underline{X})^{-1}\underline{x}' & \underline{e}_2 \\ & -1-\underline{x}(\underline{X}'\underline{X})^{-1}\underline{x}' & 0 \\ & & 0 \end{bmatrix}$$

must be formed, and the SWEEP operator must be applied to the diagonal elements of $I-\underline{X}_2(\underline{X}'\underline{X})^{-1}\underline{X}_2'$.

It may be of interest to note that the change in the regression coefficients when rows $(\underline{X}_2, \underline{Y}_2)$ are deleted is given by

$$\Delta\underline{b} = \underline{b}-\underline{b}^{(1)} = (\underline{X}'\underline{X})^{-1}\underline{X}_2'[I-\underline{X}_2(\underline{X}'\underline{X})^{-1}\underline{X}_2']^{-1}\underline{e}_2 \qquad (5.12)$$

after some algebra using partitioned matrices. The quantities needed for $\Delta\underline{b}$ are produced in the matrices above.

Row Deletion and Variable Augmentation - An Equivalence

Suppose that, instead of deleting the last n_2 rows from $\underset{\sim}{X}$, an expanded $\underset{\sim}{X}$ matrix is formed which will be denoted by $\underset{\sim}{X}^*$. $\underset{\sim}{X}^*$ is formed by appending to $\underset{\sim}{X}$ n_2 columns with zeros in rows $1,2,\ldots,n_1$, and an $n_2 \times n_2$ identity matrix for the last n_2 rows. i.e.,

$$\underset{\sim}{X}^* = \begin{bmatrix} \underset{\sim}{X}_1 & \underset{\sim}{0} \\ \underset{\sim}{X}_2 & I \end{bmatrix} .$$

Let also $\underline{x}^* = [\underline{x},\underline{0}]$, where $\underline{0}$ is a $1 \times n_2$ vector of zeros. Now,

$$\underset{\sim}{X}^{*\prime}\underset{\sim}{X}^* = \begin{bmatrix} \underset{\sim}{X}_1'\underset{\sim}{X}_1 + \underset{\sim}{X}_2'\underset{\sim}{X}_2 & \underset{\sim}{X}_2' \\ X_2 & I \end{bmatrix} .$$

A fundamental identity on the form of the inverse of a partitioned matrix (see for example (16), Theorem 8.2.5) yields

$$(\underset{\sim}{X}^{*\prime}\underset{\sim}{X}^*)^{-1} = \begin{bmatrix} (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1} & -(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_2' \\ -\underset{\sim}{X}_2(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1} & I + \underset{\sim}{X}_2(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_2' \end{bmatrix} .$$

Using this form of $(\underset{\sim}{X}^{*\prime}\underset{\sim}{X}^*)^{-1}$,

$$\underline{x}^*(\underset{\sim}{X}^{*\prime}\underset{\sim}{X}^*)^{-1}\underline{x}^{*\prime} = \underline{x}(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underline{x}'$$

obtains. So, obviously,

$$1+\underline{x}^*(\underline{X}^{*\prime}\underline{X}^*)^{-1}\underline{x}^{*\prime} = 1+\underline{x}(\underline{X}_1'\underline{X}_1)^{-1}\underline{x}'.$$

Consider the relationship between s_1^2 and $s^{*2} = \underline{e}^{*\prime}\underline{e}^*/(n_1-p)$, i.e., between the mean square errors for the model with the last n_2 rows deleted and the model with the new columns appended to $\underline{X}$ respectively. Clearly,

$$\underline{e}^* = \underline{Y}-\underline{X}^*(\underline{X}^{*\prime}\underline{X}^*)^{-1}\underline{X}^{*\prime}\underline{Y}$$

$$= \begin{bmatrix}\underline{Y}_1\\ \underline{Y}_2\end{bmatrix} - \begin{bmatrix}\underline{X}_1 & \underline{0}\\ \underline{X}_2 & I\end{bmatrix}\begin{bmatrix}(\underline{X}_1'\underline{X}_1)^{-1} & -(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_2'\\ -\underline{X}_2(\underline{X}_1'\underline{X}_1)^{-1} & I+\underline{X}_2(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_2'\end{bmatrix}$$

$$\begin{bmatrix}\underline{X}_1' & \underline{X}_2'\\ 0 & I\end{bmatrix}\begin{bmatrix}\underline{Y}_1\\ \underline{Y}_2\end{bmatrix}$$

$$= \begin{bmatrix}\underline{Y}_1\\ \underline{Y}_2\end{bmatrix} - \begin{bmatrix}\underline{X}_1(\underline{X}_1'\underline{X}_1)^{-1}\underline{X}_1'\underline{Y}_1\\ \underline{Y}_2\end{bmatrix} = \begin{bmatrix}\underline{e}_1^{(1)}\\ 0\end{bmatrix}.$$

So,

$$\underline{e}^{*\prime}\underline{e}^* = (\underline{e}_1^{(1)\prime},\underline{0}')(\underline{e}_1^{(1)\prime},\underline{0}')' = \underline{e}_1^{(1)\prime}\underline{e}_1^{(1)}.$$

Therefore,

$$s^{*2} = \underline{e}^{*\prime}\underline{e}^*/(n_1-p) = \underline{e}_1^{(1)\prime}\underline{e}_1^{(1)}/(n_1-p) = s_1^2.$$

The above relations suggest an algorithm for detecting outliers. If this approach is used, any determination about

which rows are outlying must be based on the estimated coefficients corresponding to the dummy variables. A significant coefficient indicates that the corresponding observation is not adequately explained by the rest of the data and it needs its own parameter. This becomes clear if one observes that

$$\underline{b}^* = (\underline{\underline{X}}^{*\prime}\underline{\underline{X}}^*)^{-1}\underline{\underline{X}}^{*\prime}\underline{Y} = \begin{bmatrix} (\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{\underline{X}}_1'\underline{Y}_1 \\ \underline{Y}_2 - \underline{\underline{X}}_2(\underline{\underline{X}}_1'\underline{\underline{X}}_1)^{-1}\underline{\underline{X}}_1'\underline{Y}_1 \end{bmatrix} = \begin{bmatrix} \underline{b}^{(1)} \\ \underline{e}_2^{(1)} \end{bmatrix} .$$

Now the last n_2 observations are fully explained by means of the coefficients $\underline{e}_2^{(1)}$. (Compare also the expression for $\underline{e}^*$ found earlier). The significance of these coefficients is, therefore, identical to the significance of the residuals $\underline{e}_2^{(1)}$. The test statistic t_2^2/n_2 is simply the usual partial F for testing whether a set of regression coefficients is zero in the presence of other explanatory variables. The F distribution can be used only if the set $(\underline{\underline{X}}_2, \underline{Y}_2)$ has been specified in advance, and not after the data have been inspected and, say, the maximum has been chosen to be tested.

CHAPTER VI

APPLICATION

In this chapter, an application of this technique in the field of management science is discussed. The performance of the W criterion is compared with that of other commonly used selection criteria as well as with that of models proposed in independent studies by other investigators. The field of application, parametric cost estimation, is receiving considerable attention (31), (33), (34). Parametric cost estimation is a widely used method of obtaining single valued predictions of the cost of a new item, such as a weapon system. It deals with predicting the cost (response variable) of a system by means of explanatory variables (predictors) such as system characteristics or performance requirements. This procedure is based on the premise that the cost of a system is related in a quantifiable way to the system's physical and performance characteristics. The expression of this quantifiable relationship is in the form of an estimating equation derived through statistical regression analysis of historical cost data on systems which are, more or less, analogous to the proposed system. Recent experience in weapon system acquisition programs has underscored the differences between cost estimates and realized costs. This has given impetus to the search for better cost estimating techniques.

Consider the case of predicting the cost of a new aircraft based on its planned physical and performance characteristics, and the costs and characteristics of aircraft built in the past. The role of analogy is obvious in this situation. Which historical aircraft and which variables should be used? The Mahalanobis distance seems well suited to answer these questions. Any variable selection technique which ignores the issue of analogy, and which fails to give some consideration to dimensions (variables) along which there is a marked dissimilarity between the proposed system and the historical data, may lead to gross errors due to extrapolation. The W criterion has the potential of bringing this issue to the attention of the analyst, and should be used as part of a thorough investigation. In what follows, optimal models under different criteria are "automatically" computed and used in an effort to compare on a fair (or equally unfair) basis the relative performance of the W criterion. It should be underscored that this is done partly because of considerations of mathematical convenience and it may not, in all instances, agree with good practice.

The data base is given in Table I. It consists of 23 observations on 12 physical and performance characteristics of different single engine jet fighter aircraft built over an interval spanning the years from 1947 to 1969. The response variable, Y, is the flyaway unit cost (in 1972 $100,000) of the hundredth aircraft built for each type.

TABLE I. Aircraft Data

A/C	Date	Cost	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80	3/47	2.76	50.5	6.3	1.5	4.8	1.7	9	432	8176	4950	40000	1145	4500
FH-1	1949	4.76	36.8	6.0	1.5	16.5	1.9	12	421	6699	4800	34500	790	3120
F2H-1	1949	8.75	47.9	5.9	1.5	10.5	1.5	12	514	9663	8030	49600	1215	6300
F7U-1	1949	7.78	36.0	3.0	1.8	12.0	1.2	27	602	12837	15100	50000	1420	9800
F-84E	5/49	6.18	57.0	5.0	1.7	12.0	1.8	18	432	10205	6060	44200	1694	4900
F3D-1	11/50	9.50	53.1	6.3	1.6	12.0	2.0	24	415	14890	3690	32000	1145	6500
F-86H	1/53	5.14	68.0	5.6	1.5	11.5	1.6	6	600	13836	12900	49650	1573	8920
F9F-8	7/54	4.76	51.0	3.6	1.5	12.0	1.2	12	575	11628	6630	44700	1120	14500
F4D-1	1/55	16.70	34.0	2.0	1.4	5.8	1.4	60	647	15225	20150	56600	915	14800
F3H-1N	4/55	27.68	57.0	2.9	1.5	6.5	2.3	40	620	18691	3950	44000	980	10900
F-102A	6/55	26.64	37.0	2.2	1.5	4.0	1.9	36	677	19350	18700	51800	1294	16000
F-100D	1/56	13.71	78.0	3.9	1.7	6.5	2.6	26	757	20638	18100	48800	1712	16000
FJ-4	2/56	12.31	51.0	4.5	1.4	6.0	2.0	13	607	12843	9650	49000	1330	7800
F-104A	3/56	15.73	161.0	4.3	1.8	3.4	9.7	40	1037	13384	58928	64680	1398	17900
F11F-1	3/57	13.59	74.0	4.0	1.4	6.0	2.9	11	654	13307	16300	49000	1146	10500
F-105B	6/57	51.90	91.0	3.2	1.6	4.6	4.3	42	1195	29855	33800	48100	1935	24500
F-101C	9/57	20.78	109.0	4.3	1.9	6.5	4.3	60	856	29277	32900	40000	1845	30000
F-106B	10/58	29.82	51.0	2.4	1.5	4.0	2.6	38	1153	24651	42900	51800	1487	24500
F-4B	11/63	32.78	74.0	2.8	1.7	4.7	3.7	76	1186	28539	45200	54950	1528	34000
F-5A	10/64	10.12	68.0	3.9	1.7	4.8	3.3	16	505	8085	28700	50600	1145	8160
F-4J	4/67	27.84	79.0	2.8	1.7	4.7	3.9	76	1220	30328	41300	54800	1401	35800
F-111A	10/67	107.10	92.0	1.6	2.6+	3.6	4.1	120	1262	46172	26600	57050	3156	37000
F-8E	10.69	11.19	56.0	3.4	1.5	5.5	2.5	20	984	17836	22300	52350	1238	19600

+Correct value is 1.6. Value 2.6 was used by other investigators. Used in this study also for comparison purposes.

The variables, denoted by $X_1, X_2, \ldots, X_{12}$, are the values of the following characteristics:

X_1 = Wing Loading Ratio

X_2 = Aspect Ratio

X_3 = Full to Empty Weight Ratio

X_4 = Thickness-to-Cord Ratio

X_5 = Lift to Drag Ratio

X_6 = Total Avionics Input Power in kva

X_7 = Maximum Speed in knots (Clean, Combat Weight)

X_8 = Weight Empty in lbs

X_9 = Rate of Climb in ft/min, (sea level, combat weight and power)

X_{10} = Combat Ceiling in feet

X_{11} = Ferry Range in nautical miles

X_{12} = Sea Level Static Thrust (max) in lbs.

(For detailed methodologies of data determination and term definitions see (31)).

Each observation was set aside, the models found optimal under six criteria were computed based on the remaining 22 observations and the deleted row was predicted. Various statistics were also output such as M and MSE for each model. The criteria compared were:

1. Minimum MSE
2. Minimum C_k
3. Maximum F
4. R^2 employing a subjective "elbow rule"
5. MSEP
6. Minimum W (nominal 95% prediction interval).

For the R^2 criterion, the optimal model for each subset size was output. The rule employed for the final selection was: Use the model with largest R^2 whose size is such that the next largest size does not increase R^2 by more than one percent. Of the six criteria, the maximum F was the most parsimonious. It selected variable X_8 in all cases and it consistently outperformed the others in terms of the size of the absolute error of prediction. The minimum W criterion did, on the average, worse than the others. The case of the F-111A aircraft is worth mentioning, as it clearly represents a situation which warrants special attention. Its non-analogousness to the rest of the data shows up clearly in every sort of residual analysis. Historically, each military flight component needed a new aircraft. The Air Force needed a new interceptor, the Navy wanted a carrier launched attack aircraft and the Marines required an aircraft capable of ground support missions. The then Secretary of Defense, Robert McNamara, decided to have one aircraft built that would meet all requirements thereby achieving tremendous savings. The tri-service design resulted in the F-111A which, at the time, was the most sophisticated, fastest, heaviest and costly single engine jet ever built. Its design included a radical wing which could swing forward or backward depending on desired flight characteristics. (Incidentally, the F-111A experienced all kinds of technical problems, was not well received by the three services and is often referred to as

"McNamara's Second Edsel".) Its cost was underestimated dramatically by all models. Due to the fact that its weight does not conform to its other characteristics (as conformity is defined by the other aircraft), the Mahalanobis distance associated with any model which included weight as a predictor was excessively large. For this reason, weight was not included in the minimum W model, in spite of its general importance (it showed up in every model that was encountered). The error of prediction associated with this model was far above that of every other model. However, when weight was forced into the regression, the new minimum W model clearly outperformed all others. There was only a 3% increase in W, which was still much smaller than the W's of models selected by other criteria. This point will be discussed again in what follows.

All criteria performed poorly on the untransformed data in comparison to models suggested by others (Columbia Research Corporation (31), Clemson University graduate student projects in Math 805 [unpublished]) after a complete analysis. This suggested the need for further investigation of some of the models for signs of misspecification. The residuals were analyzed for the models which were encountered most frequently. In all cases, there were clear indications of gross violations of the assumptions on the errors. The residuals exhibited clear patterns when plotted against the X's and against the sorted fitted values. Plots of Y versus individual X's showed lack of

linearity which, although not necessary for the linearity of the multi-variable model, tended to confirm information in other plots. The scatter of Y versus X_8 though was more or less linear. These clear indications of model misspecification tend to explain the better performance of the most parsimonious criterion and the poor showing of the minimum W criterion. As mentioned earlier, the maximum F criterion always selected variable X_8 which is both a significant variable and linearly related to Y. The poor performance of the minimum W criterion is explained as follows:

This criterion is based on a statistic (prediction interval width) the proper interpretation of which is based on the usual assumptions on the errors. When those assumptions are grossly violated, the Mahalanobis distance may no longer be a reliable measure of analog. This is the second point to be considered when this criterion is employed. Appropriate transformations on the variables must be performed before the selection is made. Also, after the selection, an analysis of the resulting residuals is necessary in order to validate the assumptions for the selected models. Only models which seem to satisfy those assumptions should be compared by means of this criterion.

The problem of appropriate transformations is not a simple one. Theory and common sense may suggest answers to this question but in unstructured situations a thorough investigation is usually called for. There are many possibilities and a thorough investigation is needed. As a

step in making the problem feasible, all variables were transformed to their natural logarithms. This transformation is often considered appropriate for cost data (33) and, in fact, was employed in all the previously cited competing models (31). The same preliminary plots and residual analyses aluded to above indicated that models selected after this transformation did not exhibit signs of gross misspecification. The same selection procedure was used. The minimum W criterion was based on the logarithmic units. Point predictions and prediction errors in the original units were also calculated and compared. This was done by applying the exponential transformation to the point predictions. The fact that this approach is known to produce biased estimates of the conditional mean (14) should not affect the comparative value of the estimates. Transformation of prediction intervals into the original units is not easily defined so as to make comparisons meaningful. This was not needed and was not attempted. Table II shows, for each aircraft and each model, the errors of prediction in both units. For each aircraft and for each model different from the minimum W model, the "gain" was also calculated, as defined by the difference of the absolute error of that model from the absolute error of the minimum W model. Gains are shown in Table III. (The entries of Tables II and III have been rounded to two decimal places because of space considerations.) A positive gain indicates a better prediction by the minimum W model. A single zero

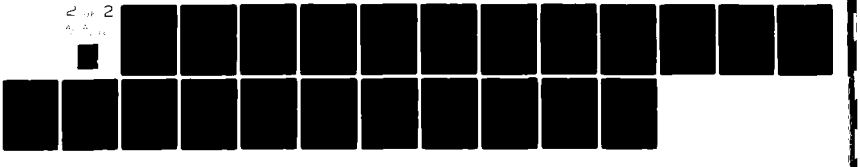
TABLE II. Absolute Errors in Logarithmic and Original Units

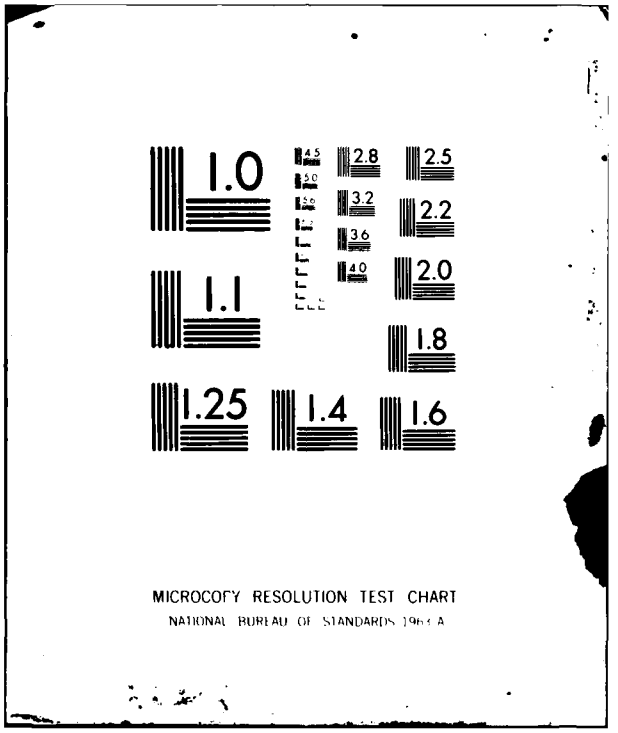
	Min MSE		Min C_k		Max F		R^2		MSEP		Min W	
A/C	Log.	Orig.	Log.	Orig.	Log.	Orig.	Log.	Orig.	Log.	Orig.	Log.	Orig.
F-80	1.06	* 5.21	1.49	* 9.44	1.49	* 9.44	1.49	* 9.44	1.04	5.07	0.52	1.89
FH-1	0.05	0.26	0.06	0.27	0.32	* 1.29	0.06	0.27	0.12	0.63	0.05	0.26
F2H-1	0.99	5.49	0.89	5.15	0.85	5.01	0.83	4.92	0.97	5.43	0.89	5.15
F7U-1	0.09	0.73	0.24	2.08	0.26	2.26	0.24	2.08	0.17	1.43	0.24	2.08
F-84E	0.82	7.82	0.27	1.95	0.13	0.89	0.27	1.95	0.64	5.50	0.27	1.95
F3D-1	0.09	0.78	0.09	0.77	0.28	3.07	0.07	0.77	0.19	2.00	0.09	0.78
F-86H	0.32	1.90	0.25	1.51	0.81	6.40	0.25	1.51	0.52	3.48	0.26	1.53
F9F-8	0.05	* 0.26	0.04	* 0.19	0.62	4.10	0.04	* 0.19	0.04	* 0.19	0.59	3.84
F4D-1	0.11	1.94	0.00	0.00	0.28	4.02	0.00	0.00	0.00	0.00	0.00	0.00
F3H-1N	0.55	11.66	0.12	3.11	0.48	10.52	0.12	3.11	0.37	8.64	0.12	3.11
F-102A	0.74	*29.14	0.02	0.45	0.38	8.50	0.10	2.44	0.46	15.39	0.10	2.44
F-100D	0.17	2.52	0.19	2.89	0.42	7.18	0.19	2.89	0.14	2.09	0.17	2.52
FJ-4	0.18	2.03	0.26	2.83	0.23	2.52	0.26	2.83	0.01	0.05	0.26	2.83
F-104A	0.71	*16.20	0.38	* 7.34	0.42	5.40	0.38	* 7.34	0.84	20.53	0.47	5.91
F11F-1	0.02	0.22	0.10	1.27	0.28	3.27	0.10	1.27	0.09	1.16	0.10	1.27
F-105B	0.45	18.94	0.39	16.59	0.42	17.77	0.39	16.50	0.45	18.87	0.39	16.59
F-101C	0.04	0.88	0.05	1.13	0.58	16.20	0.05	1.13	0.06	1.13	0.05	1.13
F-106B	0.17	4.56	0.05	1.40	0.12	3.32	0.05	1.40	0.18	4.84	0.05	1.40
F-4B	0.12	3.56	0.13	3.97	0.02	0.73	0.13	3.97	0.09	2.85	0.12	3.46
F-5A	0.61	* 4.64	0.45	* 3.64	0.84	5.73	0.45	* 3.64	0.66	4.90	0.45	3.64
F-4J	0.19	5.86	0.17	5.03	0.31	10.23	0.17	5.03	0.10	2.90	0.19	5.85
F-111A	1.09	*71.02	0.22	*25.82	0.58	*46.90	0.30	*27.52	0.80	*59.12	0.22	25.82
F-8E	0.24	2.99	0.18	2.26	0.40	5.45	0.18	2.26	0.25	3.22	0.23	2.95

TABLE III. Gain in Logarithmic and Original Units

	Min MSE		Min C_k		Max F		R^2		MSEP	
A/C	Log.	Orig.	Log.	Orig.	Log.	Orig.	Log.	Orig.	Log.	Orig.
F-80	0.54 *	3.32	0.97 *	7.56	0.97 *	7.56	0.97 *	7.56	0.52	3.18
FH-1	0	0	+0.00	0.01	0.26 *	1.03	+0.00	0.01	0.07	0.37
F2H-1	0.10	0.34	0	0	-0.04	-0.14	-0.06	-0.21	0.08	0.28
F7U-1	-0.15	-1.05	0	0	0.02	0.18	0	0	-0.07	-0.65
F-84E	0.54	5.87	0	0	-0.14	-1.06	0	0	0.36	3.56
F3D-1	0	0	-0.00	-0.01	0.19	2.29	-0.00	-0.01	0.11	1.22
F-86H	0.05	0.37	-0.00	-0.02	0.55	4.87	-0.00	-0.02	0.26	1.95
F9F-8	-0.54 *	-3.58	-0.55 *	-3.64	0.03	0.26	-0.55 *	-3.64	-0.55 *	-3.64
F4D-1	0.11	1.94	0	0	0.28	4.02	0	0	+0.00	+0.00
F3H-1N	0.43	8.56	0	0	0.35	7.41	0	0	0.26	5.53
F-102A	0.64 *	26.70	-0.08	-1.99	0.29	6.06	0	0	0.36	12.95
F-100D	0	0	0.02	0.36	0.25	4.65	0.02	0.36	-0.03	-0.43
FJ-4	-0.08	-0.79	0	0	-0.03	-0.31	0	0	-0.26	-2.78
F-104A	0.24 *	10.29	-0.09 *	1.43	-0.05	-0.51	-0.09 *	1.43	0.36	14.62
F11F-1	-0.08	-1.05	0	0	0.18	2.00	0	0	-0.01	-0.11
F-105B	0.07	2.35	0	0	0.03	1.18	0	0	0.07	2.29
F-101C	-0.01	-0.26	0	0	0.53	15.17	0	0	+0.00	+0.00
F-106B	0.12	3.16	0	0	0.07	1.92	0	0	0.13	3.44
F-4B	0	0	0.01	0.41	-0.09	-2.83	0.01	0.41	-0.02	-0.71
F-5A	0.17 *	1.01	0 *	0	0.39	2.09	0 *	0	0.22	1.26
F-4J	0	0	-0.02	-0.83	0.12	4.37	-0.03	-0.83	-0.09	-2.96
F-111A	0.87 *	45.20	0 *	0	0.36 *	21.07	0.08 *	1.70	0.59 *	33.30
F-8E	+0.00	0.04	-0.05	-0.69	0.16	2.50	-0.05	-0.69	0.02	0.27

AD-A097 535 CLEMSON UNIV SC DEPT OF MATHEMATICAL SCIENCES F/G 21/1
AN EMPIRICAL MODEL BUILDING CRITERION BASED ON PREDICTION WITH --ETC(U)
AUG 80 A S KORKOTSIDES, K T WALLENIUS N00014-75-C-0451
UNCLASSIFIED N117 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

indicates that the model selected coincided with the minimum W model. The gains are shown for both original and logarithmic units. An asterisk in Tables II and III indicates that the aircraft under prediction was at a large Mahalanobis distance along the variables selected by the corresponding criterion. For each criterion, Table IV shows the average absolute error and the average percent absolute error in logarithmic units. The average percent absolute error is the average absolute error as a percentage of the observed response value. The average percent gain is defined similarly and it is also shown, together with the sum of gains, in Table IV for the five other criteria. Table V shows the corresponding statistics in the original units. From these two tables, it can be seen that the minimum W criterion outperformed all others on all counts on the average over the 23 observations. The minimum C_k criterion did comparably well and, as can be observed from Table III, it selected the same model as the minimum W criterion more often than any other. In view of its relation to the problem of prediction discussed in Chapter II, this should not be surprising. Notice also that this criterion did better than the minimum W criterion more often than not, although the difference in some cases was very small. The F104-A aircraft was predicted better by the minimum W model in the original units. The performance of the R^2 criterion was also comparable to that of the minimum W criterion, however, a definite statement cannot be made due to the fact that

TABLE IV. Performance Statistics in Logarithmic Units

	Min MSR	Min C_k	Max F	R^2	MSEP	Min W
Average |err|	0.385	0.262	0.456	0.266	0.356	0.253
Avrg. % |err|	17.33	14.43	22.70	14.50	16.78	12.15
Total Gain	3.03	0.21	4.68	0.31	2.37	
Avrg. % Gain	5.18	2.28	10.55	2.34	4.62	

TABLE V. Performance Statistics in Original Units

	Min MSR	Min C_k	Max F	R^2	MSEP	Min W
Average |err|	8.636	4.308	7.838	4.459	7.366	4.195
Avrg. % |err|	42.13	33.55	54.33	33.84	39.66	25.15
Total Gain	102.15	2.59	83.78	6.06	72.92	
Avrg. % Gain	17.15	8.40	29.18	8.69	14.50	

TABLE VI. Average Nominal 95% Prediction Interval Widths and Frequency of Coverage in Logarithmic Units

	Min MSR	Min C_k	Max F	R^2	MSEP	Min W
Avrg. Width	1.385	1.250	1.778	1.287	2.288	1.223
Coverage	86.96	91.30	91.30	91.30	95.65	95.65

the R^2 model is not objectively defined. In fact, if a more parsimonious rule had been employed, the performance of this criterion would have deteriorated.

As was mentioned earlier, the minimum W statistic will tend to underestimate the true 95% prediction interval. The average interval was calculated for each criterion, as well as the percentage of observations which were actually covered by the corresponding prediction intervals. These are shown in Table VI. It is a pleasant surprise that, in spite of the observation above, the prediction intervals associated with the minimum W models covered the observed responses more often than the others. The occurrence of this phenomenon in the problem at hand may not provide firm ground on which a claim that it will happen in general can be based. Nevertheless, it seems that the prediction intervals associated with the minimum W models are well centered about the expected value of y. This is a very desirable property.

Referring to Table III, the observations which the minimum W model failed to predict well were examined further. This was done in an effort to identify common features which might serve as a warning in a careful analysis. The F9F-8 is the only case in which the W criterion was definitely outperformed by four of the other criteria. The model selected consisted of variables X_2, X_5, X_8 and X_9. This model was never selected for any other observation by any criterion. More importantly, variable X_{12} was conspicuously

absent. This variable was present in all the models with the smaller prediction errors for each observation, and it would seem desirable to force it into the final predictive equation. When this was done, i.e., when the model with smallest W among those containing this variable was selected, the resulting errors of prediction in logarithmic and in original units were 0.056 and 0.259 respectively, a definite improvement. The new model, although it represented a 30% increase in W, still had the smallest W of all models selected by the other criteria.

In the case of the F-104A the situation is not as clear. The minimum W criterion again failed to select variable X_{12}. However, its error in logarithmic units was not much greater than the errors of the minimum C_k, maximum F and R^2 criteria which did better and, in fact, in original units the error was smaller than that of the minimum C_k and R^2 models. A closer investigation revealed that the Mahalanobis distance of this observation from the variables in the minimum MSE, minimum C_k and R^2 models was very large. In contrast, in the case of the F9F-8, the reduction in the Mahalanobis distance attained by the omission of variable X_{12} was not as dramatic. When this variable was forced into the minimum W model for the F-104A, the logarithmic error was the smallest observed, (-0.35) although the error in the original units became slightly larger (-6.68).

The above observations suggest the need for a careful examination of such cases. It is as important that

extrapolation be avoided, as it is that important variables be retained. The way in which these two considerations should be balanced against each other is a matter of judgment on the part of the analyst. The minimum W criterion has the potential for calling attention to such issues.

The case of the F-80 aircraft is the one clearly favoring this criterion. Considerable reduction in the Mahalanobis distance was attained, without worsening the fit, by the simple switching of certain variables while retaining the important ones, namely X_8 and X_{12}. (As in the case of the untransformed data, variable X_8 was contained in all of the better models. It was always selected by all criteria except MSEP.

The models selected by each criterion and the corresponding observations are given in Tables VII through XII. For each criterion, the models selected were ranked according to their frequency of occurrence and their ranks were used for labeling purposes. The models selected by the MSEP criterion are all marked by "×". Observe that twenty different models were selected by this criterion, none of which was ever selected by any other criterion, due to the fact that they were all associated with small R^2 values. In view of the fact (mentioned in Chapter II) that this criterion ignores the MSE (and every other measure of fit) of the postulated submodels, this should not be surprising. Observe also the large average prediction interval.

TABLE VII. Minimum W Models and Aircraft Predicted

A/C	Variables											
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80		2	2		2			2				2
FH-1		2	2		2			2				2
F2H-1	3	3	3	3	3		3	3	3		3	
F7U-1		1			1			1				1
F-84E		1			1			1				1
F3D-1		2	2		2			2				2
F-86H		2	2		2			2				2
F9F-8		4			4			4	4			
F4D-1		1			1			1				1
F3H-1N		1			1			1				1
F-102A		1			1			1				1
F-100D		2	2		2			2				2
FJ-4		1			1			1				1
F-104A		5						5				
F11F-1		1			1			1				1
F-105B		1			1			1				1
F-101C		1			1			1				1
F-106B		1			1			1				1
F-4B		2	2		2			2				2
F-5A		2	2		2			2				2
F-4J		2	2		2			2				2
F-111A		1			1			1				1
F-8E		2	2		2			2				2

TABLE VIII. Minimum MSE Models and Aircraft Predicted

A/C	Variables											
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80		3		3	3			3				3
FH-1		1	1		1			1				1
F2H-1	6	6	6	6	6		6	6	6		6	
F7U-1		1	1		1			1				1
F-84E	5	5	5	5	5	5	5		5		5	
F3D-1		1	1		1			1				1
F-86H	2	2	2	2	2		2	2	2	2	2	
F9F-8		1	1		1			1				1
F4D-1		1	1		1			1				1
F3H-1N	7	7	7	7	7		7	7		7	7	
F-102A	8	8	8	8	8		8	8	8	8	8	8
F-100D		1	1		1			1				1
FJ-4		1	1		1			1				1
F-104A	2	2	2	2	2		2	2	2	2	2	
F11F-1		1	1		1			1				1
F-105B	2	2	2	2	2		2	2	2	2	2	
F-101C		1	1		1			1				1
F-106B	2	2	2	2	2		2	2	2	2	2	
F-4B		1	1		1			1				1
F-5A	2	2	2	2	2		2	2	2	2	2	
F-4J		1	1		1			1				1
F-111A	4	4	4	4	4	4	4	4			4	
F-8E	2	2	2	2	2		2	2	2	2	2	

TABLE IX. Minimum C_k Models and Aircraft Predicted

A/C	Variables											
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80				3		3		3				3
FH-1		1			1			1				1
F2H-1	4	4	4	4	4		4	4	4		4	
F7U-1		1			1			1				1
F-84E		1			1			1				1
F3D-1		1			1			1				1
F-86H		1			1			1				1
F9F-8		1			1			1				1
F4D-1		1			1			1				1
F3H-1N		1			1			1				1
F-102A		2	2		2			2				2
F-100D		1			1			1				1
FJ-4		1			1			1				1
F-104A		1			1			1				1
F11F-1		1			1			1				1
F-105B		1			1			1				1
F-101C		1			1			1				1
F-106B		1			1			1				1
F-4B		1			1			1				1
F-5A		2	2		2			2				2
F-4J		1			1			1				1
F-111A		1			1			1				1
F-8E		1			1			1				1

TABLE X. Maximum F Models and Aircraft Predicted

A/C	Variables											
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80				2		2		2				2
FH-1								1				
F2H-1		3	3		3			3				3
F7U-1								1				
F-84E								1				
F3D-1								1				
F-86H								1				
F9F-8								1				
F4D-1								1				
F3H-1N								1				
F-102A								1				
F-100D								1				
FJ-4								1				
F-104A								1				
F11F-1								1				
F-105B								1				
F-101C								1				
F-106B								1				
F-4B								1				
F-5A								1				
F-4J								1				
F-111A								1				
F-8E								1				

TABLE XI. R^2 Models and Aircraft Predicted

A/C	Variables											
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80				3		3		3				3
FH-1		1			1			1				1
F2H-1		1			1			1				1
F7U-1		1			1			1				1
F-84E		1			1			1				1
F3D-1		1			1			1				1
F-86H		1			1			1				1
F9F-8		1			1			1				1
F4D-1		1			1			1				1
F3H-1N		1			1			1				1
F-102A		1			1			1				1
F-100D		1			1			1				1
FJ-4		1			1			1				1
F-104A		1			1			1				1
F11F-1		1			1			1				1
F-105B		1			1			1				1
F-101C		1			1			1				1
F-106B		1			1			1				1
F-4B		1			1			1				1
F-5A		2	2		2			2				2
F-4J		1			1			1				1
F-111A		2	2		2			2				2
F-8E		1			1			1				1

TABLE XII. MSEP Models and Aircraft Predicted

A/C	Variables											
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
F-80	×		×		×						×	
FH-1		×			×							
F2H-1	×	×	×	×	×	×	×	×	×			
F7U-1							×	×			×	×
F-84E										×		
F3D-1					×							
F-86H			×		×				×			×
F9F-8	×		×		×							
F4D-1				×					×			
F3H-1N						×		×				
F-102A		×		×						×		
F-100D					×		×			×		
FJ-4					×					×	×	
F-104A				×			×					
F11F-1		×							×			
F-105B								×			×	
F-101C							×					
F-106B					×			×				
F-4B	×	×	×		×						×	
F-5A						×		×				
F-4J		×			×							
F-111A					×	×						×
F-8E					×							

With respect to the minimum W models, it seems that most early and later aircraft were predicted by one model, while most of the ones in the middle of the time scale were predicted by another. Three aircraft required their own models. The F2H-1 was not predicted well by any criterion and it was the one observation whose deletion from the data base reduced dramatically the widths of the prediction intervals associated with the minimum W model for all but one of the other observations. The studentized residual (discussed in Chapter V) which was associated with this aircraft was very large (4.2 on the average) for each and every aircraft under prediction. The F9F-8 was selected for deletion in the one remaining case. (In general, the performance of the minimum W criterion improved when one observation was deleted for each prediction. A detailed exposition is not given since the deletion of observations is not advocated in this dissertation.) The one observation whose deletion reduced W the most in each case, and the percent reduction in the width of the "prediction interval" are shown in Table XIII. The new "prediction intervals" contained the observed y's only 86.96% of the time.

Finally, it should be emphasized that the purpose of this analysis being the gaining of insight into the relative performance of the minimum W criterion, certain aspects of the problem (which a complete analysis should not fail to consider) where not stressed. Data determination and model form specification, for instance, received only secondary attention.

TABLE XIII. Observation Deleted and Maximum Reduction in Width of the Prediction Interval for Each Aircraft

A/C	A/C Deleted	Reduction (%)
F-80	F2H-1	33.23
FH-1	F2H-1	31.06
F2H-1	F-102A	26.01
F7U-1	F2H-1	25.20
F-84E	F2H-1	25.17
F3D-1	F2H-1	33.22
F-86H	F2H-1	16.39
F4D-1	F2H-1	25.48
F3H-1N	F2H-1	25.59
F-102A	F2H-1	25.51
F-100D	F2H-1	31.31
FJ-4	F2H-1	29.53
F-104A	F9F-8	7.37
F11F-1	F2H-1	26.08
F-105B	F2H-1	28.79
F-101C	F2H-1	25.20
F-106B	F2H-1	25.49
F-5A	F2H-1	39.37
F-4J	F2H-1	32.78
F-111A	F2H-1	29.64
F-8E	F2H-1	31.78

Other data sets were also analyzed in less detail. A limited simulation study was conducted, in which the correlation matrix, the number of variables and the number of observations were allowed to vary. Although a complete investigation would be a large project and was not attempted, the observations made during these studies seem to support the ones made on the aircraft data. In the absence of variables which appeared fundamental to the predictive equation, the minimum W criterion consistently outperformed the other criteria whenever it succeeded in reducing a large Mahalanobis distance. This was more pronounced in the cases where the correlation structure involved high multicollinearity. In these last cases, large Mahalanobis distances were frequently reduced significantly by the exclusion of variables which caused the multicollinearity.

Although in no way conclusive, it may be of interest to note that the minimum W models performed better (along the same lines discussed above) than models suggested by other investigators after careful analyses on the aircraft data. This in no way means that mechanical selection of variables is preferable to a careful investigation. The minimum W criterion should be used as part of a complete analysis. As is the case with virtually every data analytic technique, pedestrian application can result in curious and misleading conclusions. There is no substitute for a careful, reasoned analysis.

CHAPTER VII

DISCUSSION AND CONCLUSIONS

This investigation has been concerned with the problem of predicting the response at a known point in the space of the explanatory variables in the context of multiple linear regression. This is the problem with which parametric cost estimation is concerned and it is encountered frequently in other applications. The view taken in this dissertation is that the issue of analogy of the point under prediction to the historical data should not, in such cases, be ignored.

The Mahalanobis distance has been studied as an appropriate measure of analog in higher dimensions. The width of the prediction interval is a numeraire which combines this measure of analog with the mean square error, which is a standard measure of the fit provided by a given model. Thus, the prediction interval offers itself as a tool with which variables can be screened and models brought in the foreground that are reasonable candidates for the purpose of such analyses. The experience gained by applying this methodology to real and simulated data sets suggests that the careful analyst should benefit from its use by gaining insight on an aspect of the problem which otherwise would not have been brought into focus. As was mentioned earlier, the careful analyst would certainly become skeptical about a

model which, although otherwise reasonable, produced a very large prediction interval at the point under consideration. However, the W criterion has the advantage of providing a clear and unconfused warning by taking the issue of analogy into consideration during the screening process. The W criterion, just as any other, should not be used in a pedestrian way as a method for pointing to "the one best model". The notion of a model which is best for all purposes is not defined in unstructured situations which comprise the bulk of empirical model building. Even for a specific application, a claim about the knowledge of such a model can not be defended on uncontestable grounds. Therefore, selection criteria ought to be viewed as screening aids and used as such. This point, although generally accepted, is all too frequently forgotten in practice.

There is a second point on which the W criterion may prove to be a valuable aid. The exclusion of variables which are known to be important from other considerations ought to be taken as a warning about the peculiarity of the point under prediction. Selection criteria may occasionally point to models which the analyst finds unacceptable either because of the variables which they contain (or exclude) or because of the fact that the underlying model assumptions seem to be violated. In such cases, the usual statistics lose their validity and risks attendant with the use of a suspect model are introduced. If no reasonable model can be

found which passes model aptness considerations, the regression approach to the problem based on the given body of data should be questioned.

In parametric cost estimation where the cost of an object system must be predicted based on historical data on "similar" systems built in the past, the notion of analogy often presents itself conspicuously. It is conceivable that the proposed system may reflect, in the values of the explanatory variables associated with it, technological developments and/or performance characteristics different from those encountered in the historical data along certain dimensions. A new weapon system, for instance, would probably not even be considered if such were not the case. A model which fails to explain the historical data adequately would be unreasonable to use for the prediction of the new system. It seems equally unreasonable, however, to devote all effort into fitting the historical data, disregarding the relation of the proposed system to them. The W criterion can (and should) be employed, together with other considerations, so that both aspects of the problem will be given deserved attention if the final model is not to be grossly myopic.

Models which are found good by more than one criteria are highly desirable. The W criterion can be employed to suggest several models which can then be compared with those suggested by other criteria. This procedure will focus the

attention of the analyst on a (hopefully) small number of models which must be carefully scrutinized before a final choice is made.

The distributional properties of W under selection pose a highly complex problem which has not been investigated in this dissertation. Another problem which has not been considered is the following: The statistic W is expressed in the units of the response variable used in the selection process. Often, various transformations on the response variable are considered in the same problem. How is one to compare the W's associated with models based on different transformations on the response variable? This is a question which can only be answered on a case by case basis. In some cases it may be a simple mathematical problem, while in others it may defy an objective definition.

Finally, although an extensive simulation study was not conducted, such a study may be a worthwhile endeavor that can provide useful insight into the questions raised in this investigation.

BIBLIOGRAPHY

1. Allen, D. M. "Mean Square Error of Prediction as A Criterion for Selecting Variables." Technometrics 13: 469-476. 1971.

2. Allen, D. M. "The Relationship between Variable Selection and Data Augmentation and A Method for Prediction." Technometrics 16: 125-127. 1974.

3. Allen, D. M. "Comment." Journal of the American Statistical Association 72: 95-96. 1977.

4. Bingham, Christopher. "Some Identities Useful in the Analysis of Residuals from Linear Regression." Unpublished Technical Report. University of Minnesota. 1977.

5. Box, G. E. P. "Use and Abuse of Regression." Technometrics 8: 625-629. 1966.

6. Chambers, M. J. "Regression Updating." Journal of the American Statistical Association 66: 744-748. 1971.

7. Cook, R. D. "Detection of Influential Observations in Linear Regression." Technometrics 19: 15-18. 1977.

8. Cook, R. D. "Influential Observations in Linear Regression." Journal of the American Statistical Association 74: 169-174. 1979.

9. Daniel, C., and F. S. Wood. Fitting Equations to Data, John Wiley & Sons, Inc. New York. 1971.

10. Draper, N. R., and H. Smith. Applied Regression Analysis, John Wiley & Sons, Inc. New York. 1966.

11. Edwards, T. B. "On the Degree of Inflation of Measures of Fit Induced by Empirical Model Building." Unpublished Ph.D. Dissertation. Department of Mathematical Sciences. Clemson University. 1979.

12. Furnival, M. G., and R. W. Wilson, Jr. "Regressions by Leaps and Bounds." Technometrics 16: 499-510. 1974.

13. Gentleman, F. J., and M. B. Wilk. "Detecting Outliers II. Supplementing the Direct Analysis of Residuals." Technometrics 31: 387-410. 1975.

14. Goldberger, S. A. Topics in Regression Analysis, The Macmillan Company. New York. 1968.

15. Gorman, W. J., and R. J. Toman. "Selection of Variables for Fitting Equations to Data." Technometrics 8: 27-51. 1966.

16. Graybill, A. F. Theory and Application of the Linear Model, Duxbury Press. North Scituate, Massachusetts. 1976.

17. Harris, J. R. A Primer of Multivariate Statistics, Academic Press. New York. 1975.

18. Hoaglin, C. D., and R. E. Welsch. "The Hat Matrix in Regression and ANOVA." The American Statistician 32: 17-22. 1978.

19. Hocking, R. R. "The Analysis and Selection of Variables in Linear Regression." Biometrics 32: 1-49. 1976.

20. Hotelling, H. "The Generalization of Student's Ratio." Presented at the meeting of the American Mathematical Society at Berkeley, April 11, 1931.

21. Johnston, J. Econometric Methods, (2nd ed.), McGraw-Hill. New York. 1972.

22. Leamer, E. E. Specification Searches, John Wiley & Sons, Inc. New York. 1978.

23. Lindley, D. V. "The Choice of Variables in Multiple Regression." Journal of Royal Statistical Society 30: 31-53. 1968.

24. Mahalanobis, P. C. "On Tests and Measures of Group Divergence." Journal of Asiatic Society (Bengal) 26: 541-588. 1930.

25. Mallows, C. L. "Choosing Variables in A Linear Regression: A Graphical Aid." Presented at the Central Region Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas. 1964.

26. Mallows, C. L. "Some Comments on C_p." Technometrics 15: 661-675. 1973.

27. Mason, R. L., J. T. Webster, and R. F. Gunst. "Sources of Multicollinearity in Regression Analysis." Communications in Statistics 4: 277-292. 1975.

28. Mosteller, F., and J. W. Tukey. Data Analysis and Regression, Addison-Wesley Publishing Co. 1977.

29. Ralston, A., and H. S. Wilf, (eds). Mathematical Methods for Digital Computers, John Wiley & Sons, Inc. New York. 1960.

30. Rao, C. R. Linear Statistical Inference and Its Applications, John Wiley & Sons, Inc. New York. 1965.

31. Sponsler, G. C., D. Gignoux, and N. N. Rubin. "Parametric Cost Estimation of Fighter Aircraft." Technical Report. Columbia Research Corporation. Gaithersburg, Md. 1973.

32. Theil, H. Economic Forecasts and Policy, North Holland Publishing Co., Amsterdam. 1961.

33. Timson, F. S., and D. P. Tihansky. "Confidence in Estimated Airframe Costs: Uncertainty Assessment in Aggregate Predictions." Technical Report. Rand Corporation. Santa Monica, Ca. 1972.

34. Wallenius, K. T. "Mahalanobis Distance as A Measure of Analog in Parametric Cost Estimation." Unpublished Technical Report. Department of Mathematical Sciences, Clemson University. 1978.

35. Walls, C. R., and D. L. Weeks. "A Note on The Variance of a Predicted Response in Regression." The American Statistician 23: 24-26. 1969.

36. Welsch, E. R., and S. C. Peters. "Finding Influential Subsets of Data in Regression Models." Center for Computational Research. Sloan School of Management, Massachusetts Institute of Technology. 1979.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER N117	2. GOVT ACCESSION NO. AD-A097537	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) An Empirical Model Building Criterion Based on Prediction and Applications in Parametric Cost Estimation		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER TR #349
7. AUTHOR(s) A. S. Korkotsides K. T. Wallenius		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS Clemson University Dept. of Mathematical Sciences Clemson, South Carolina 29631		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 047-202
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Code 434 Arlington, Va. 22217		12. REPORT DATE 8-8-80 13. NUMBER OF PAGES 110
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multiple Regression, prediction, empirical model building, Parametric Cost Estimation.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In the context of multiple linear regression, when a subset of k-out-of-p predictor variables is to be selected for the purpose of predicting the response at some known point in the predictor variables' space, the width of the resulting prediction interval gives an indication of the precision with which the response is predicted and, thus, it may provide a suitable selection criterion.

(OVER)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. (continued)

A review of commonly used selection criteria is given, with special emphasis on those which deal with the problem of prediction. The Mahalanobis distance is one of the quantities affecting the width of the prediction interval, and it is studied in some detail. The effects of adding a new variable to a model are investigated and a monotonicity theorem is derived.

The influence of an observation on the width of the prediction interval, as measured by the effected change when that observation is set aside, is also investigated and an equivalence between observation deletions and variable augmentation is shown.

The relationships found in these investigations indicate the applicability of certain computing techniques. Computing algorithms are presented.

A management science application os the statistical procedures developed in this study is explored in the area of parametric cost estimation.